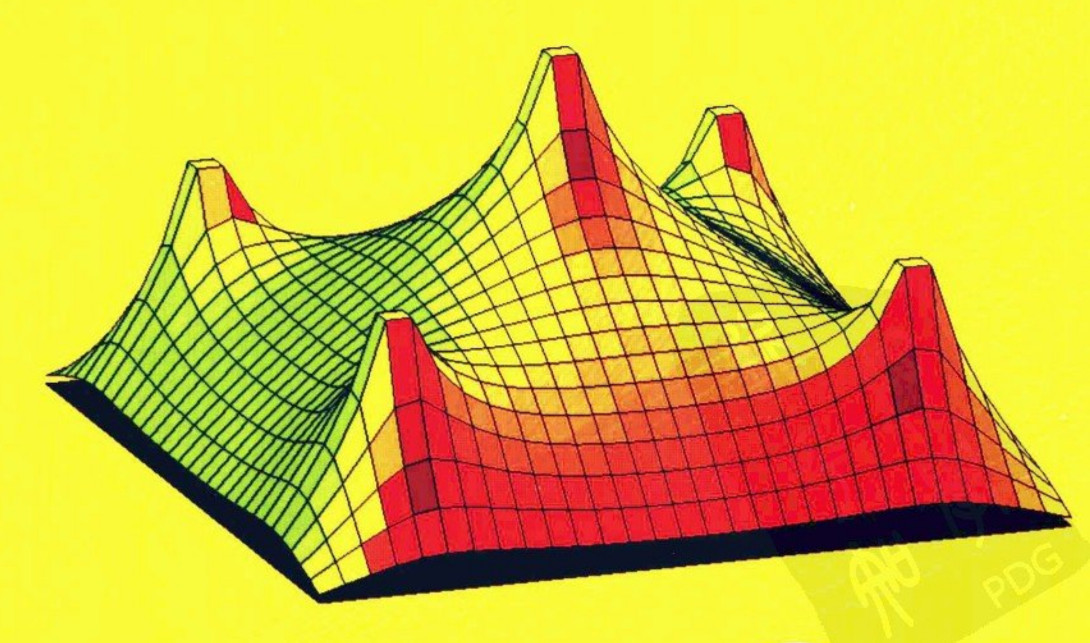
- 合理、完善的知识体系结构
- 内容丰富，重点突出，应用性强
- 免费提供相关程序源代码下载
- 深入、详细剖析 MATLAB 工程应用技术

MATLAB 工程应用书库

MATLAB 概率与数理统计分析



张德丰　等编著

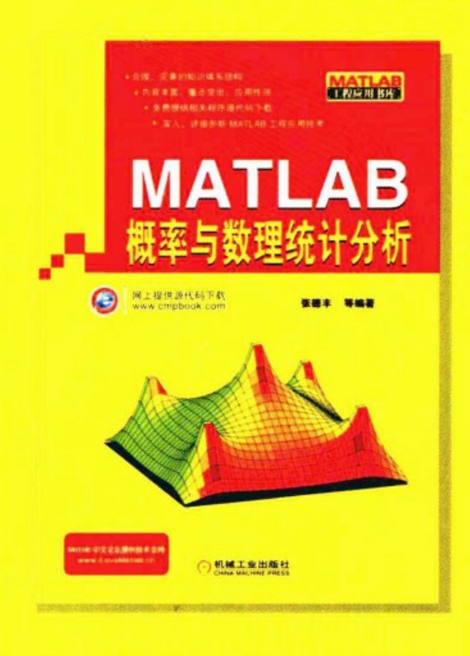




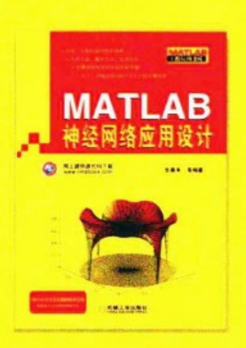
Matlab 中文论坛提供技术支持
www.iLoveMatlab.cn



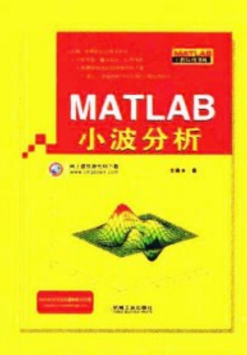
ISBN 978-7-111-29325-5
策　划：丁诚　吴鸣飞
封面设计：ZSC 子时文化 ZiShi Culture



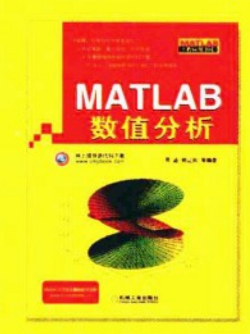
ISBN 978-7-111-29325-5
定价：41.00元



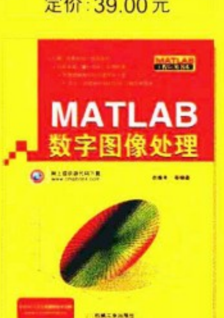
ISBN 978-7-111-25612-0
定价：39.00元



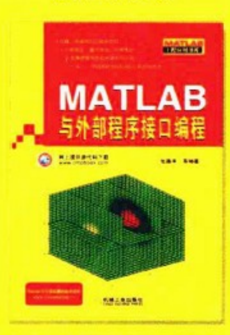
ISBN 978-7-111-25613-7
定价：41.00元



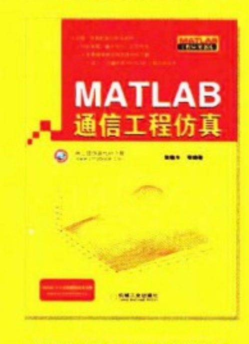
ISBN 978-7-111-25707-3
定价：39.00元



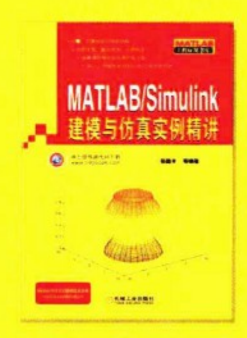
ISBN 978-7-111-25735-6
定价：39.00元



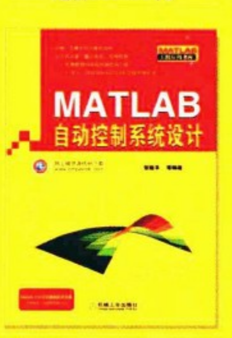
ISBN 978-7-111-25706-6
定价：42.00元



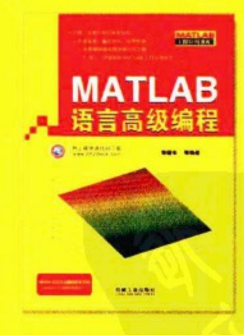
ISBN 978-7-111-29323-1
定价：46.00元



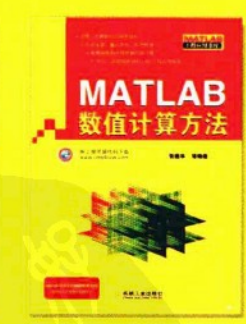
ISBN 978-7-111-29326-2
定价：46.00元



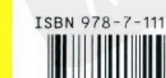
ISBN 978-7-111-29308-8
定价：45.00元



ISBN 978-7-111-29265-4
定价：45.00元



ISBN 978-7-111-29324-8
定价：42.00元





地址：北京市百万庄大街22号　邮政编码：100037
电话服务　网络服务
社服务中心：(010)88361066　门户网：http://www.cmpbook.com
销售一部：(010)68326294　教材网：http://www.cmpedu.com
销售二部：(010)88379649
读者服务部：(010)68993821　封面无防伪标均为盗版

定价：41.00元

MATLAB概率与数理统计分析

张德丰　等编著



机 械 工 业 出 版 社

本书采用最新版 MATLAB R2009a，介绍概率与统计的基本原理、典型应用，以及使用 MATLAB 进行实际工程中概率与统计分析的基本方法。本书共分 9 章。第 1 章介绍 MATLAB 的数据基础，第 2 章介绍概率与数理统计基本概念，第 3 章介绍多维随机向量，第 4 章介绍统计估计及统计特征，第 5 章介绍统计检验方法——假设检验，第 6 章介绍方差分析及曲线拟合，第 7 章介绍回归分析，第 8 章介绍多元统计分析，第 9 章介绍隐马尔可夫模型及统计工具箱的示范程序等内容。

本书可作为工科硕士研究生应用概率与统计课程的教材和非数学与统计类专业本科高年级学生的选修教材，也可作为管理、科研和工程技术人员的参考用书。

图书在版编目（CIP）数据

MATLAB 概率与数理统计分析 / 张德丰等编著. —北京：机械工业出版社，2010.1

（MATLAB 工程应用书库）

ISBN 978-7-111-29325-5

Ⅰ. M…　Ⅱ. 张…　Ⅲ. ①概率论－统计分析－计算机辅助计算－软件包，MATLAB②数理统计－统计分析－计算机辅助计算－软件包，MATLAB Ⅳ. O21-39

中国版本图书馆 CIP 数据核字（2009）第 233732 号

机械工业出版社（北京市百万庄大街 22 号　邮政编码 100037）

策划编辑：丁　诚　吴鸣飞

责任编辑：丁　诚　吴超莉

责任印制：洪汉军

三河市宏达印刷有限公司印刷

2010 年 1 月第 1 版 · 第 1 次印刷

184mm × 260mm · 22.5 印张 · 557 千字

0001 – 4000 册

标准书号：ISBN 978-7-111-29325-5

定价：41.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换

电话服务

社服务中心：（010）88361066

销 售 一 部：（010）68326294

销 售 二 部：（010）88379649

读者服务部：（010）68993821

网络服务

门户网：http://www.cmpbook.com

教材网：http://www.cmpedu.com

前　　言



MATLAB 是一种主要用于工程计算的高级计算机语言。美国的 MathWorks 公司自 1984 年推出 MATLAB 的 DIS 版本后，又推出了它的 Windows 版本，并且不断推出更新的版本，使得 MATLAB 的应用领域越来越广。到目前为止，MATLAB 已经包括仿真工具 Simulink、自动控制、信号处理、图像处理、神经网络、模式识别、小波分析、数理统计、生物信息等 30 多个工具箱。由于其灵活的编程方法和极高的编程效率，加上其在用户界面和功能上的不断扩展，自推出以来，日益受到广大高校师生和科研人员的青睐。

MATLAB R2009a 是 MATLAB 的新版本，它对以往版本中的产品模块进行了一些调整。例如，MATLAB Builder for COM 的功能被集成到了 MATLAB Builder for .NET 中，Finacial Time Series Toolbox 的功能被集成到了 Financial Toolbox 中。MATLAB 将高性能的数值计算和可视化集成在一起，并提供了大量的内置函数，从而被广泛地应用于科学计算、控制系统、信息处理等领域的分析、仿真和设计工作。利用 MATLAB 产品的开放式结构，可以非常容易地对其功能进行扩充，从而不断深化对工程问题的认识。

MATLAB 开放的产品体系使其成为了诸多领域开发的首选软件。MATLAB 还具有 500 余家第三方合作伙伴，分布在科学计算、机械动力、化工、计算机通信、汽车和金融等领域。接口方式包括了联合建模、数据共享和开发流程衔接等。

为了更好地适应高等院校培养高等技术应用型人才的需要，解决高等院校“概率与统计”理论课与实践课相结合的问题，并根据应用数学与专业相融、基础数学为专业服务的基本要求和以应用为目的、以必需与够用为度的基本原则，作者在多年从事高等教育教学实践的基础上，编写了本书。

本书介绍概率与数理统计的基本原理、典型应用，以及使用 MATLAB 进行实际工程分析的基本方法。全书共分 9 章。第 1 章介绍 MATLAB 的数据基础；第 2 章介绍概率与数理统计基本概念，包括随机事件及其概率、事件及运算、条件概率与事件的独立性等内容；第 3 章介绍多维随机向量，包括二维随机向量、随机向量的分布、二维随机向量的数字特征等内容；第 4 章介绍统计估计及统计特征，包括统计图的绘制、变量分布估计及概率分布的统计特征等内容；第 5 章介绍统计检验方法——假设检验，包括假设检验概述、单正态总体的假设检验等内容；第 6 章介绍方差分析及曲线拟合，包括因素方差分析及数据曲线拟合等内容；第 7 章介绍回归分析，包括一元线性回归分析、多元线性回归分析、偏最小二乘回归分析等内容；第 8 章介绍多元统计分析，包括因素分析、聚类分析及判别分析等内容；第 9 章介绍隐马尔可夫模型及统计工具箱的示范程序，包括隐马尔可夫模型、示范程序等内容。

本书具有如下特点：

第一，注意概率与数理统计的思想方法介绍。在阐述某一统计概念方法时，一般是从具体实例开始引出相关内容，或是以具体实例结束相关内容。

第二，本书在重视公式和定理推导的同时，也重视 MATLAB 应用于概率与数理统计方法时的简单性、实用性和可操作性。实际中，概率与统计几乎遍及各个领域，成为解决实际

问题的重要工具。

第三，突出了知识的技能化和应用意识的养成。

通过对本书的学习，读者不仅可以掌握概率与数理统计的内容，同时也能初步掌握使用MATLAB 进行数据处理的基本方法和技巧。

参加本书编写的有张德丰、许华兴、王旭宝、王孟群、邓恒奋、卢国伟、卢焕斌、伍志聪、庄文华、庄浩杰、许业成、何沛彬、何佩贤、张水兰、张坚、李勇杰、李秋兰、李美妍、陈运英、陈景棠、梁家科、黄达中、陈楚明、林健锋、梁劲强、林振满、周品。

由于作者水平有限，书中难免存在不足之处，敬请读者批评指正。

作　者





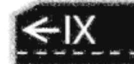
目　录













第 1 章　MATLAB 的数据基础

MATLAB 代表 Matrix Laboratory，是一个高性能的科学计算平台，集成了数值计算、矩阵计算和图形绘制等众多功能。在 MATLAB 中，问题的提出和解答只需按一般的数学方式表达和描述，不需要大量原始而传统的编程过程，因此它特别适用于研究、解决工程和数学问题。MATLAB 还具有易扩展性，每个使用者都可以自定义编写函数或程序。

1.1　MATLAB 的主要功能

1.1.1　MATLAB 简介

启动 MATLAB 后，系统将自动打开命令窗口，如图 1-1 所示。

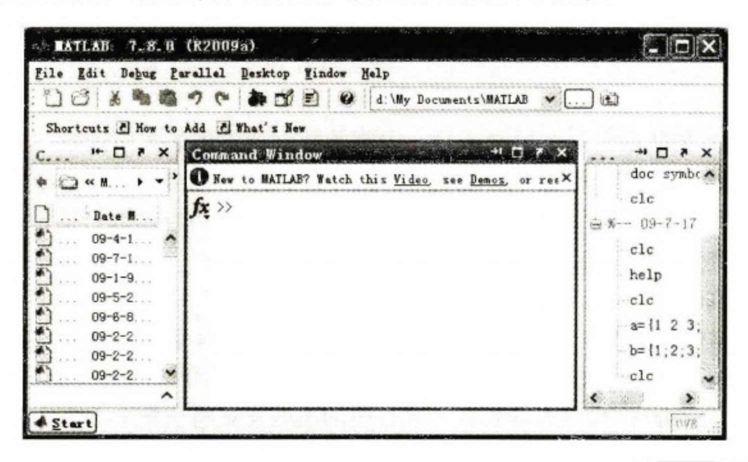


图 1-1　MATLAB 命令窗口

MATLAB 按默认的路径启动，可以通过 pathtool 函数改变路径，即在命令窗口中进行“>>pathtool”即可打开路径设置对话框。图中“>>”为 MATLAB 命令的提示符，显示正在等待执行命令。此时，如果输入相应的命令，MATLAB 就会运行，并得到运动结果。可以使用光标键〈→〉、〈↑〉、〈↓〉或〈→〉调用前面的命令，以及在命令行中移动光标位置以修改命令。

单击【File】菜单下【New】子菜单下的【M-file】选项，或单击工具栏中的 按钮，则弹出程序编辑窗口，如图 1-2 所示。

MATLAB 的变量、注释与标点、函数及 Script 文件介绍如下。

1. 变量

MATLAB 变量的命名应遵守一定的规则：变量以字母开头（区分大小写），之后可以是



任意字母、数字或下画线，但最长不能超过 36 个字符，也不能与 MATLAB 中的特殊变量（如 ans、pi、eps、inf、NaN、i、j、nargin、nargout 等）同名。

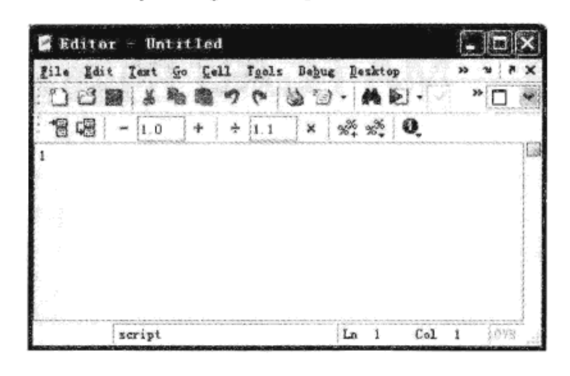


图 1-2　MATLAB 程序编辑窗口

2．注释与标点

命令行中“%”符号后的所有文字为注释，计算机不会执行。多条命令可以放在同一行，但要用逗号或分号隔开。命令后的逗号表示显示结果，分号则禁止显示结果。

符号“…”表示语句的余下部分将出现在下一行，但它不能出现在变量名或运算符之间。

3．函数

MATLAB 自带的函数称为内部函数。内部函数支持常用的数学函数，书写方式也基本一致，如三角函数 sin(*x*)、常用对数 lg(*x*)、自然对数 log(*x*)、指数函数 exp(*x*)，开方根 sqrt(*x*)等。用户还可以编写各种自定义的函数，然后像 MATLAB 内部函数一样，在工作环境下、Script 文件和其他函数中调用。

4．Script 文件

将 MATLAB 命令放在一个文件中，然后告诉 MATLAB 打开文件并顺次执行其中的命令，这个文件被称为 Script 文件。它可通过单击【File】菜单下【New】子菜单下的【M-file】选项创建。Script 文件具有全局性，文件中的所有变量将在整个工作环境中有效。

Script 文件可直接在编辑或工作窗口中执行，也可被其他 M 文件和函数调用。在工作窗口直接输入 Script 文件名便可运行，而在编辑窗口中运行 Script 文件需单击【Debug】菜单下的【Save File and Run】命令，然后切换到工作窗口查看运行结果。

1.1.2 MATLAB 的数据及数值分析

1．数据分析

MATLAB 在作数据分析时，如果输入的是向量，运算是对整个向量进行的；若输入的是数组（矩阵），则运算按列进行。

利用 MATLAB 可进行数据的基本统计计算，如下列各种函数。运算时，如果调用格式中有 dim，则指明运算按指定维数进行。

1）max(x, dim)：求最大元素。

2）min(x, dim)：求最小元素。

3）median(x, dim)：求中位值。

4）mean(x, dim)：求平均值。

5）std(x, flag)：求标准差，flag 指明标准差的不同计算方式。

6）prod(x, dim)：求积。



7）sum(x, dim)：求和。

8）cumsum(x, dim)：求累计和。

9）cumprod(x, dim)：求累计积。

10）cov(x)：求协方差阵。

11）cov(x, y)：求相关阵。

12）corrcoef(x)：求两随机变量的协方差。

13）corrcoef(x, y)：求两随机变量的相关系数。

14）sort(x)：以升序排列元素。

2. 微积分的分析

1）limit：函数的极限。

其调用格式如下：

```
limit(F,x,a)
limit(F,a)
limit(F)
limit(F,x,a,'right')
limit(F,x,a,'left')
```

其中，limit(F,x,a)：计算符号表达式 $F=F(x)$ 的极限值，当 $x \to a$ 时；limit(F,a)：计算函数 F 的极限，当 $x=a$ 时；limit(F)：默认 $a=0$ 时，求函数 F 的极限；limit(F,x,a,'right')：计算符号函数 F 的右极限，当 $x \to a^{+}$ 时；limit(F,x,a,'left')：计算符号函数 F 的左极限，当 $x \to a^{-}$ 时。

【例 1-1】 求函数极限示例。

```
>> syms x a t h;
a1=limit(sin(x)/x)
a2=limit(1/x,x,0,'right')
a3=limit(1/x,x,0,'left')
a4=limit((sin(x+h)-sin(x))/h,h,0)
v = [(1 + a/x)^x, exp(-x)];
a5=limit(v,x,inf,'left')
```

运行程序，输出如下：

```
a1 =   1
a2 = Inf
a3 = -Inf
a4 = cos(x)
a5 = [ exp(a),         0]
```

2）fminbnd：求单变量函数的极值。

其调用格式如下：

```
x=fminbnd(fun,x1,x2)
```

其中，x = fminbnd(fun,x1,x2)：计算在区间 $a-b$ 上函数 F 取最小值时的 x 值。

【例 1-2】 求函数 $f(x)=2x^3-6x^2-18x+7$ 在区间 $(-2,4)$ 的极小值，并作图。

其实现的 MATLAB 程序代码如下：

```
f=inline('2*x.^3-6*x.^2-18*x+7');          %建立内联函数 f(x)
[x,fl]=fminbnd(f, -2,4)                    %求函数 f 的最小值和对应的 x 值
fplot(f,[-2,4]);                           %作图
```

运行程序，输出如下（效果见图 1-3）：

```
x =      3.0000
fl =   -47.0000
```

注意：用 inline 建立的函数 f，在 fminbnd 和 fplot 命令中不用加单引号，而用 M 函数文件建立的函数则要加单引号。

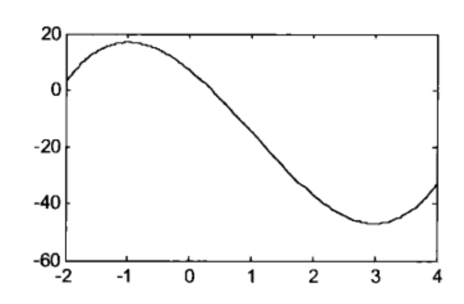


图 1-3　fminbnd 函数图形结果

3）diff：求函数的微分。

其调用格式如下：

```
Y = diff(X,n,dim)
```

其中，Y = diff(X,n,dim)：对符号表达式 X 中指定的符号变量 n 计算 X 的 dim 阶导数。在默认状态下，s=findsym(X)，dim=1。

【例 1-3】 求函数的微分示例。

```
>> syms x y t
d1=diff(sin(x^2)*y^2,2)      %计算 ∂²y²/∂x² sin x²
d2=diff(d1,y)                %计算 ∂/∂y(∂²y²/∂x² sin x²)
d3=diff(t^6,6)
```

运行程序，输出如下：

```
d1 =
-4*sin(x^2)*x^2*y^2+2*cos(x^2)*y^2
d2 =
-8*sin(x^2)*x^2*y+4*cos(x^2)*y
d3 = 720
```



4）quad：求函数的积分。

其调用格式如下：

```
q = quad(fun,a,b)
q = quad(fun,a,b,tol)
[q,fcnt] = quad(...)
```

其中，q = quad(fun,a,b)：近似地从 a 到 b 计算函数 fun 的数值积分，误差为10^{-6}；q = quad(fun,a,b,tol)：用指定的绝对误差 tol 代替默认误差；[q,fcnt] = quad(...)：返回函数 fcnt 的估计阶数。

【例 1-4】 求积分函数 $y=\int_0^2 \frac{1}{x^3-2x-5}\mathrm{d}x$。

其实现的 MATLAB 程序代码如下：

```
F = @(x)1./(x.^3-2*x-5);
Q = quad(F,0,2);
```

运行程序，输出如下：

```
Q =     -0.4605
```

5）trapz：梯形法数值积分。

其调用格式如下：

```
Z = trapz(X,Y)
```

其中，Z = trapz(X,Y)：用梯形法计算 Y 在 X 点上的积分。

【例 1-5】 用梯形法求数值积分示例。

```
X = sort(rand(1,101)*pi);
Y = sin(X);
Z = trapz(X,Y)
```

运行程序，输出如下：

```
Z =      1.9989
```

6）int：符号函数的积分。

其调用格式如下：

```
R=int(S,v)
```

```
R=int(S,v,a,b)
```

其中，R=int(S,v)：对符号表达式 S 中指定的符号变量 v 计算不定积分；R=int(S,v,a,b)：对表达式 S 中指定的符号变量 v 计算从 a 到 b 的定积分。

【例 1-6】 符号函数的积分示例。

```
syms x z t alpha
R1=int(-2*x/(1+x^2)^2)
R2=int(x/(1+z^2),z)
R3=int(x*log(1+x),0,1)
R4=int(2*x, sin(t), 1)
R5=int([exp(t),exp(alpha*t)])
```

运行程序，输出如下：

```
R1 = 1/(1+x^2)
R2 = x*atan(z)
R3 = 1/4
R4 = 1-sin(t)^2
R5 = [exp(t), 1/alpha*exp(alpha*t)]
```

7）taylor：泰勒级数展开。

其调用格式如下：

```
taylor(f)
taylor(f,n,v,a)
```

其中，taylor(f)：求出符号函数 f 在 x=0 处的 5 阶麦克劳林型泰勒展开式；taylor(f,n,v,a)：求出符号函数 f 在 v=a 点的 n–1 阶泰勒展开式。

【例 1-7】 求二阶泰勒级数展开示例。

```
syms a x
f=a/(x-10);
y1=taylor(f,x,3)          %求 f 在 x=0 处的二阶泰勒级数展开
y2=taylor(f,3,x,4)        %求 f 在 x=4 处的二阶泰勒级数展开
```

运行程序，输出如下：

```
y1 =
-1/10*a-1/100*a*x-1/1000*a*x^2
y2 =
-1/6*a-1/36*a*(x-4)-1/216*a*(x-4)^2
```

8）傅里叶级数展开。

MATLAB 中没有专门用于傅里叶级数展开的命令，可编写一个 M 文件实现傅里叶级数展开。

```
function [a0,an,bn]=mfourier(f)
```

```
syms n x
a0=int(f,-pi,pi)/pi;
an=int(f*cos(n*x),-pi,pi)/pi;
bn=int(f*sin(n*x),-pi,pi)/pi;
```



【例 1-8】 傅里叶级数展开示例。

```
syms x
f=x^2+x;
[a0,an,bn]=mfourier(f)
```

运行程序，输出如下：

```
a0 = 2/3*pi^2
an =
2*(-2*sin(pi*n)+n^2*sin(pi*n)*pi^2+2*n*cos(pi*n)*pi)/n^3/pi
bn =
-2/n^2*(-sin(pi*n)+n*cos(pi*n)*pi)/pi
```

进一步化简：

```
>> an=simple(an)
an = -4/n^3/pi*sin(pi*n)+2/n*sin(pi*n)*pi+4/n^2*cos(pi*n)
>> bn=simple(bn)
bn = 2/n^2/pi*sin(pi*n)-2/n*cos(pi*n)
```

3. 非线性方程的数值解

1）fsolve：最小二乘法。

其调用格式为

```
x = fsolve(fun,x0)
```

其中，x = fsolve(fun,x0)：求方程 *fun*=0 在估计值 *x*0 附近的近似解。

【例 1-9】 求方程 $x-\mathrm{e}^{-x}=0$ 的解。

```
fc=inline('x-exp(-x)');
x1=fsolve(fc,0)
```

运行程序，输出如下：

```
Optimization terminated: first-order optimality is less than options.TolFun.
x1 =
     0.5671
```

【例 1-10】 求解下列方程组的解。

$$\begin{cases}2x_1-x_2-\mathrm{e}^{-x_1}=0\\-x_1+2x_2-\mathrm{e}^{-x_2}=0\end{cases}$$

先编制函数 myfun.m 文件。



```
function F = myfun(x)
F = [2*x(1) - x(2) - exp(-x(1));
      -x(1) + 2*x(2) - exp(-x(2))];
```

在命令窗口调用 myfun 文件实现程序：

```
x0 = [-5; -5];
options=optimset('Display','iter');
[x,fval] = fsolve(@myfun,x0,options)
```

运行程序，输出如下：

```
fter 33 function evaluations, a zero is found.
                                        Norm of    First-order  Trust-region
Iteration Func-count      f(x)          step       optimality    radius
     0         3       23535.6                      2.29e+004       1
     1         6       6001.72              1       5.75e+003       1
     2         9       1573.51              1       1.47e+003       1
     3        12       427.226              1             388       1
     4        15       119.763              1             107       1
     5        18       33.5206              1            30.8       1
     6        21       8.35208              1            9.05       1
     7        24       1.21394              1            2.26       1
     8        27      0.016329       0.759511           0.206     2.5
     9        30   3.51575e-006      0.111927         0.00294     2.5
    10        33   1.64763e-013    0.00169132       6.36e-007     2.5
Optimization terminated successfully:
 First-order optimality is less than options.TolFun
x =
    0.5671
    0.5671
fval =
  1.0e-006 *
     -0.4059
     -0.4059
```

2）fzero 函数：零点法。

其调用格式如下：

```
x = fzero(fun,x0)
```

其中，x = fzero(fun,x0)：求函数 *fun* 在 *x*0 附近的零点。估计值 *x*0 若为标量，则在 *x*0 附近查找零点；*x*0=[*x*1, *x*2]为向量时，则首先要满足函数 fun(*x*1)fun(*x*2)<0，然后将严格在[*x*1, *x*2]区间内寻找零点，若找不到，系统将给出提示。

【例 1-11】 求函数 $f(x)=x^3-2x-5$ 的零点。

```
f = @(x)x.^3-2*x-5;
z = fzero(f,2)
```

运行程序，输出如下：

```
z =     2.0946
```

4．solve 函数



功能：求代数方程的符号解。

其调用格式如下：

```
solve(eq)
solve(eq,var)
solve(eq1,eq2,...,eqn)
g = solve(eq1,eq2,...,eqn,var1,var2,...,varn)
```

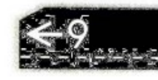
其中，solve(eq)：求解方程 eq=0，输入参量 eq 可以是符号表达式或字符串表达式；solve(eq,var)：对 eq 中指定的变量 var 求解方程 eq(var)=0；solve(eq1,eq2,...,eqn)：求解方程组 eq1=0，eq2=0，…，eq*n*=0；g = solve(eq1,eq2,...,eqn,var1,var2,...,varn)：对方程组 eq1, eq2,…,eq*n* 中指定的 *n* 个变量 var1, var2,…, var*n* 求解。

下面通过程序代码来了解 Sdve 函数的用法。

```
solve('a*x^2 + b*x + c')
solve('a*x^2 + b*x + c','b')
S = solve('x + y = 1','x - 11*y = 5')
A = solve('a*u^2 + v^2', 'u - v = 1', 'a^2 - 5*a + 6')
y1=A.a,y2=A.u,y3=A.v
```

运行程序，输出如下：

```
ans =
 -1/2*(b-(b^2-4*a*c)^(1/2))/a
 -1/2*(b+(b^2-4*a*c)^(1/2))/a
ans = -(a*x^2+c)/x
S =     x: [1x1 sym]
        y: [1x1 sym]
A =
        a: [4x1 sym]
        u: [4x1 sym]
        v: [4x1 sym]
y1 =
  2
  2
  3
  3
y2 =
 1/3+1/3*i*2^(1/2)
 1/3-1/3*i*2^(1/2)
 1/4+1/4*i*3^(1/2)
 1/4-1/4*i*3^(1/2)
```



```
y3 =
  -2/3+1/3*i*2^(1/2)
  -2/3-1/3*i*2^(1/2)
  -3/4+1/4*i*3^(1/2)
  -3/4-1/4*i*3^(1/2)
```

5. solver 函数

功能：求常微分方程的数值解。

其调用格式如下：

```
[T, Y]=solver(odefun, tspan, y0)
```

其中，[T, Y]=solver(odefun, tspan, y0)：在区间 tspan=[*t*0 *tf*]上，用初始条件 *y*0 求解显式微分方程 $y'=f(t,y)$ 。solver 为命令 ode45，ode23，ode113，ode15s，ode23s，ode23t，ode23tb 之一。

odefun 为显式常微分方程 $y'=f(t,y)$ 。

tspan 积分区间（即求解区间）的向量 tspan=[*t*0, *tf*]。要获得问题在其他指定时间点 *t*0, *t*1, *t*2,⋯,*tf* 上的解，则令 tspan=[*t*0, *t*1, *t*2,⋯, *tf*]（要求是单调的）。

*y*0 包含初始条件的向量。

求解具体 ODE 的基本过程如下：

① 根据问题所属学科中的规律、定律和公式，用微分方程与初始条件进行描述。

$$F(y,y',y'',\cdots,y^{(n)},t)=0$$

$$y(0)=y_0,y'(0)=y_1,\cdots,y^{(n-1)}(0)=y_{n-1}$$

而 $y=(y;y(1);y(2);\cdots;y(m-1))$ ，n 与 m 可以不等。

② 运用数学中的变量替换：$y_n=y^{(n-1)},y_{n-1}=y^{(n-2)},\cdots,y_2=y',y_1=y$ ，把高阶（大于二阶）的方程（组）写成一阶微分方程组：

$$\boldsymbol{Y}'=\begin{pmatrix}y_1'\\y_2'\\\vdots\\y_n'\end{pmatrix}=\begin{pmatrix}f_1(t,Y)\\f_2(t,Y)\\\vdots\\f_n(t,Y)\end{pmatrix},\quad \boldsymbol{Y}_0=\begin{pmatrix}y_1(0)\\y_2(0)\\\vdots\\y_n(0)\end{pmatrix}=\begin{pmatrix}y_0\\y_1\\\vdots\\y^{n-1}\end{pmatrix}$$

③ 根据①与②的结果，编写能计算导数的 M 函数文件 odefile。

④ 将文件 odefile 与初始条件传递给求解器 Solver 中的一个，运行后就可得到 ODE 的、在指定的时间区间上的列向量解 y（其中包含 y 及不同阶的导数）。

因为没有一种算法可以有效地解决所有的 ODE 问题。为此，MATLAB 提供了多种求解器 Solver。对于不同的 ODE 问题，采用不同的 Solver，见表 1-1。

表 1-1 不同求解器 Solver 的特点

求解器 Solver	ODE 类型	特 点	说 明
ode45	非刚性	一步算法，4, 5 阶 Runge-Kutta 方程，累计截断误差达 $(\Delta x)^3$	大部分场合的首选算法
ode23	非刚性	一步算法，2, 3 阶 Runge-Kutta 方程，累计截断误差达 $(\Delta x)^3$	适用于精度较低的情形
ode113	非刚性	多步法，Adams 算法，高低精度均可达到 $10^{-3}\sim10^{-6}$	计算时间比 ode45 短

（续）

求解器 Solver	ODE 类型	特　点	说　明
ode23t	适度刚性	采用梯度算法	适用于适度刚性情形
ode15s	刚性	多步法，Gear's 反向数值微分，精度中等	ode45 失效时，可尝试使用
ode23s	刚性	一步法，2 阶 Rosebrock 算法，低精度	当精度较低时，计算时间比 ode15s 短
ode23tb	刚性	即 TR-BDF2 实现，类似于 ode23s	这个算法比 ode15s 更精确



【例 1-12】 求解微分方程 $y' = -2y + 2x^2 + 2x$，$0 \leqslant x \leqslant 0.5$，$y(0) = 1$。

```
fun=inline('-2*y+2*x^2+2*x','x','y');
[x,y]=ode23(fun,[0 0.5],1);
x'
ans =
  Columns 1 through 8
     0    0.0400    0.0900    0.1400    0.1900    0.2400    0.2900    0.3400
  Columns 9 through 12
    0.3900    0.4400   0.4900     0.5000
>> y'
ans =
  Columns 1 through 8
   1.0000    0.9247    0.8434    0.7754    0.7199     0.6764     0.6440     0.6222
  Columns 9 through 12
   0.6105    0.6084     0.6154     0.6179
>> plot(x',y','o')
hold on;
plot(x',y')
```

运行程序，效果如图 1-4 所示。

【例 1-13】 求解描述振荡器的经典的 Ver der Pol 微分方程 $\frac{\mathrm{d}^2y}{\mathrm{d}t^2} - \mu(1-y^2)\frac{\mathrm{d}y}{\mathrm{d}t} + y = 0$，$y(0)=1$，$y'(0)=0$。

分析：令 $x_1 = y$，$x_2 = \frac{\mathrm{d}y}{\mathrm{d}t}$，$\mu = 7$，则：

$$\frac{\mathrm{d}x_1}{\mathrm{d}t} = x_2$$
$$\frac{\mathrm{d}x_2}{\mathrm{d}t} = 7(1-x_1^2)x_2 - x_1$$

编写 M 文件 VDP.m 如下：

```
function fy=VDP(t,x)
fy=[x(2);7*(1-x(1)^2)*x(2)-x(1)];
```

在命令窗口中执行以下程序：

```
Y0=[1;0];
[t,x]=ode45('VDP',[0 40],Y0);
y=x(:,1);
dy=x(:,2);
plot(t,y,t,dy);
```

运行程序，效果如图 1-5 所示。

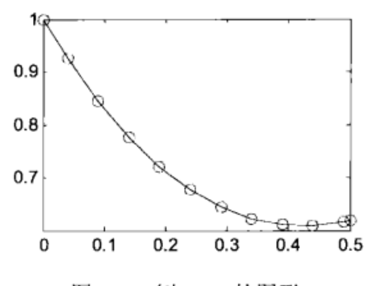


图 1-4　例 1-12 的图形

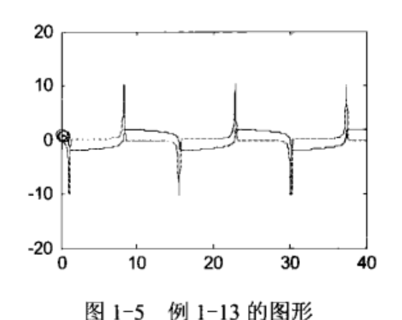


图 1-5　例 1-13 的图形

6．dsolve 函数

功能：求常微分方程的符号解。

其调用格式如下：

```
r = dsolve('eq1,eq2,...', 'cond1,cond2,...', 'v')
```

其中：

- 对给定的常微分方程（组）eq1, eq2,…中指定的符号自变量 v，与给定的边界条件和初始条件 cond1, cond2,…求符号解（即解析解）r。
- 若没有指定变量 v，则默认变量为 t；在微分方程（组）的表达式 eq 中，大写字母 D 表示对自变量（设为 x）的微分算子：$\mathrm{D}y=\mathrm{d}y/\mathrm{d}x$，$\mathrm{D}2y=\mathrm{d}^2y/\mathrm{d}x^2$，…，微分算子 D 后面的字母则表示因变量，即待求解的未知函数。
- 初始和边界条件由字符串表示：$y(a)=b$，$\mathrm{D}y(c)=d$，$\mathrm{D}2y(e)=f$ 等，分别表示 $y(x)\big|_{x=a}=b$，$y'(x)\big|_{x=c}=d$，$y''(x)\big|_{x=e}=f$。
- 若边界条件少于方程（组）的阶数，则返回的结果 r 中会出现任意常数 $C_1,C_2,\cdots$ 若该命令找不到解析解，则返回警告信息，同时返回空的 sym 对象，这时，用户可以用命令 ode23 或 ode45 求解方程组的数值解。

【例 1-14】 dsolve 函数示例。

```
D1=dsolve('Dx = -a*x')
D2=dsolve('(Dy)^2 + y^2 = 1','s')
D3=dsolve('Dy = a*y', 'y(0) = b')
D4=dsolve('D2y = -a^2*y', 'y(0) = 1', 'Dy(pi/a) = 0')
D5=dsolve('Dx = y', 'Dy = -x')
```

运行程序，输出如下：

```
D1 = C1*exp(-a*t)
D2 =
               -1
   sin(s-C1)
                1
  -sin(s-C1)
D3 =   b*exp(a*t)
D4 =   cos(a*t)
D5 =
      x: [1x1 sym]
      y: [1x1 sym]
```



1.1.3 MATLAB 矩阵的建立及基本操作

1. 数、数组、矩阵的输入

（1）数的输入

```
>> a=7
```

输出如下：

```
a =
       7
```

（2）输入复数

```
>> b=3+3i
```

输出如下：

```
b =
    3.0000 + 3.0000i
```

（3）数组的输入

```
>> X=[3 5 7;2 8 9;11 22 32]   %行之间用分号或空格隔开
```

输出如下：

```
X =
       3      5      7
       2      8      9
      11     22     32
```

（4）等待键盘的输入命令

其调用格式为

```
>> n=input('请输入初始量,n=');
```

输出如下：



```
请输入初始量,n=
```

注意：变量名开头必须是英文字母，后面的字符可以是英文、数字或下画线，但不包含空格和标点。

2. 矩阵大小的测试和定位

```
>> A=[3 5 9;2 8 9;7 8 2;1 2 3];
>> d=numel(A)                 %测试矩阵 A 的元素数
>> [n,m]=size(A)              %测试 A 的行(n)、列(m)数
>> [i,j]=find(A>3);           %找出 A 中大于 3 的元素所在的行数和列数
>> i',j'                      %输出 i，j 的转置
```

输出如下：

```
d =    12
n =     4
m =     3
ans =
     3     1     2     3     1     2
ans =
     1     2     2     2     3     3
```

注意：“%”后面是注释语句，被忽略而不执行。对一个数组可用 n=length(A)，A 若是矩阵，n 给出 A 的行、列数的最大值。

3. 矩阵的块操作

```
>> A(2,:);                    %取出 A 的第二行的所有元素
A([1 3],:);                   %取出 A 的第一，三行的所有元素
A(2:3,1:2)                    %取出 A 的第二，三行与第一，二列交叉的元素
ans =
     2     8
     7     8
>> A(2,:);                    %取出 A 的第二行的所有元素
A([1 3],:);                   %取出 A 的第一，三行的所有元素
A(2:3,1:2)                    %取出 A 的第二，三行与第一，二列交叉的元素
A([1 3],:)=A([3 1],:)         %将 A 的第一行和第三行交换
ans =
     2     8
     7     8
A =
     7     8     2
     2     8     9
     3     5     9
     1     2     3
>> A(2,:)=4;                  %将 A 的第二行的所有元素用 4 取代
>> A(find(A==3))=-3;          %将 A 中等于 3 的所有元素换为-3
```

```
>> A(2,:)=[]                %删除 A 的第二行
A =
     7     8     2
    -3     5     9
     1     2    -3
diag(A,k);                  %提取矩阵 A 的第 k 条对角线上的元素
tril(A,k);                  %抽取矩阵 A 的第 k 条对角线下面的部分
triu(A,k);                  %抽取矩阵 A 的第 k 条对角线上面的部分
```



注意："："表示"全部"的意思。

4．矩阵的翻转操作

```
flipud(A);         %对 A 进行上下翻转
fliplr(A);         %对 A 进行左右翻转
rot90(A);          %对 A 逆时针旋转 90°
```

5．特殊矩阵的产生

```
A=eye(n);               %产生 n 维单位矩阵
A=ones(n,m);            %产生 n×m 维的全 1 矩阵
A=zeros(n,m);           %产生 n×m 维的全 0 矩阵
A=rand(n,m);            %产生 n×m 维随机矩阵（元素在 0~1 之间）
randn(m,n);             %产生 m×n 维正态分布随机矩阵
B=logspace(a,b,n);      %在 a,b 之间产生 n 个对数等分向量
diag(a,b,n);            %返回 n 阶以 a, b, c, d, …为对角线元素的矩阵
hilb(n);                %返回 n 阶 Hilbert 矩阵，其元素为 H(i,j)=1/(i+j-1)
magic(n);               %产生 n 阶魔方矩阵
randperm(n);            %产生 1~n 之间整数的随机排列
```

【例 1-15】 randperm 函数示例。

>> A=randperm(6)

运行程序，输出如下：

```
A =
     6     3     5     1     2     4
```

6．数的运算

```
3+6;
3*6;
6/3;          %6 右除 3,等于 2
6\3;          %6 左除 3,等于 0.5
6^3;          %6 的 3 次方
sqrt(3);      %3 的算术平方根
exp(3);       %e 的 3 次方,不能输成 e^3
log(4);       %4 的自然对数,log10(4)表示以 10 为底,log2(4)表示以 2 为底
```



7. 矩阵的运算

```
A';               %A 的转置
det(A);           %A 的行列式,A 必须是方阵
rank(A);          %A 的秩
inv(A);           %A 的逆
eig(A);           %A 的本征值
[X,D]=eig(A);     %A 的本征矢量 X 及本征值 D
trace(A);         %A 的迹,等于 A 的对角线元素之和
3*A;              %常数与矩阵相乘
A+B;              %表示矩阵 A 与矩阵 B 相加，其中，A，B 必须是同维矩阵，和 3+A 进行比较
A-B;              %表示矩阵 A 与矩阵 B 相减，其中，A，B 必须是同维矩阵，和 3–A 进行比较
A*B;              %表示矩阵 A 与矩阵 B 相乘，和 A.*B 进行比较
A/B;              %表示矩阵 A 左除矩阵 B，和 A./B 进行比较
A\B;              %表示矩阵 A 右除矩阵 B，和 A.\B 进行比较
A^2;              %A^2 相当于 A*A，和 A.^2 进行比较
```

注意：矩阵的加、减、乘、除按相关规则运算，否则给出警告信息。“.*”，“./”，“.\”，“.^”称为点运算（或称为数组运算，又称为元素群运算）。点运算是前后矩阵对应元素之间的运算。

8. 变量的存储与调用

（1）存储

```
>> save data a b c   %将变量 a, b, c 存到 data.mat 文件中
```

（2）调用

```
>> load data   %将 data.mat 文件中的所有变量加载到工作空间
```

9. 列出工作空间所有变量

```
>> whos          %将列出工作空间所有变量的变量名、大小、字节数、数组维数
```

10. 联机帮助

```
>> help sqrt     %将显示出平方根 sqrt 命令的功能和使用方式
```

1.1.4 符号运算

符号运算是 MATLAB“符号数学工具箱”具有的功能，它是指运算对象允许是非数值的符号变量。

（1）符号表达式

符号表达式是代表数字、函数、算子和变量的 MATLAB 字符串，或字符串数组，不要求变量有预先的值。

符号表达式可以直接用单引号括起来表示，也可以用 sym、syms 或 inline 命令创建。

```
%下列语句中，表示 x，y 为自变量，f 是符号表达式，但这个方法不能创建符号方程
```



```
syms x y n
'1/(2*x^n)'                              %表示单引号内是字符串
'a*x^2+b*x+c=0'                          %表示单引号内为代数方程
f=inline('x^2+5')                        %定义函数
f=sym('a*x^2+b*x+c=0')                   %表示f的符号方程式
a=sym('[2*x,sin(x);sqrt(x),cos(x)]')     %创建2×2的符号矩阵
f=sin(x)+cos(y)
```

在符号表达式中，MATLAB 约定 D 表示一阶微分，D2 表示二阶微分，…，符号 Dy 相当于 dy/d*t*，因此

```
>> f='(Dy)^2+y^2=1'
```

表示微分方程。

在符号表达式中，如果变量数多于一个，除非特别声明，否则只有一个是独立变量。一般 *x* 永远是独立变量，可以使用函数 symvar 询问自由变量。符号变量不一样，运算结果是不同的。

（2）符号表达式运算

1）符号与数值间的转换。

digits(d)：设置有效数字个数为 *d* 的近似精度。

vap(s)：返回表达式 *s* 在 digits 函数设置下的精度的数值解。

vap(s, d)：返回表达式 *s* 在 digits(d)精度下的数值解。

subs(s, old, new)：以 new 代替表达式 *s* 中的 old。其中 old 为表达式中的符号变量，new 为符号或数值变量或数值表达式。

2）数值矩阵转换为符号矩阵。常数（数值）也可以表示为符号表达式。因此，符号矩阵运算过程中如有数值矩阵，必须将其转换成符号矩阵。

```
>> a=[10.5 11;9.1 5.6];        %矩阵
>> f=sym(a)                    %说明a是符号矩阵或将其转换成了符号矩阵
f =
[  21/2,     11]
[ 91/10,   28/5]
```

（3）运算

函数符号的运算可以通过 funtoolGUI 界面进行，各种运算的意义可通过界面上的〈Help〉键获得。其中，符号的微分、差分计算可以用与数值微分、差分相同的符号 diff。

diff(s, 'x', n)：对符号表达式 *s* 中的自变量 *x* 进行 *n* 次求导（*n* 默认值为 1）。

int(f,'a', 'b')：对符号表达式 *s* 中的自变量 *x* 在[*a*, *b*]区间进行积分。

符号矩阵的运算则与数值矩阵的运算完全相同，如：

```
>> a=sym('[x+2;3*x+3]');
>> b=sym('[x^2;3*sin(x)-3]');
>> y=a+b
```

```
y =
        x+2+x^2
 3*x+3*sin(x)
```

1.1.5 MATLAB 的绘图功能

1. 绘制二维图形

（1）基本绘图函数 plot

其调用格式如下：

```
plot(X, Y)
```

其中，plot(X, Y)：以 *X*，*Y* 的对应元素为坐标绘制二维图形，其中 *X*，*Y* 的维数要匹配。

【例 1-16】 plot 函数示例 1。

```
x=0:pi/18:2*pi;      %给出横坐标
y=sin(x);            %计算出纵坐标
plot(x,y);           %绘制图形
```

运行程序，效果如图 1-6 所示。

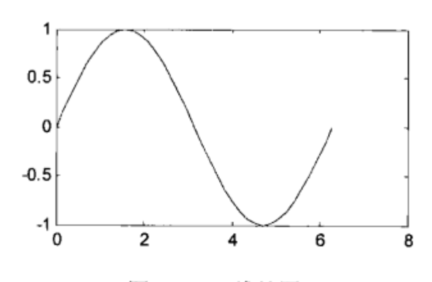


图 1-6　二维绘图

plot(Y)：若 *Y* 为 *m* 维向量，则等价于 plot(X,Y)，其中，*X*=1：*m*。

plot(X1, Y1, LineSpec1, X2, Y2, LineSpec2,…)：将按顺序分别画出由 3 个参数定义 Xi，Yi，LineSpeci（*i*=1,2,…,*n*）的线条。其中，参数 LineSpeci 指明了线条的类型，标记符号和画线的颜色。

其中，① 线型：有实线、点线、虚线、点画线。例如，“-” 表示实线。

② 线条宽度 LineWidth 取值为整数。例如，'LineWidth'，2 表示线宽为两个像素。

③ 线条颜色：常用的有 8 种颜色。例如，'b--'表示画蓝色虚线。

④ 标记类型：表示数据点标记的类型，常用的有 13 种。例如，'*r'表示红色星号。

⑤ 标记大小：MarkerSize 指定标记符号的大小尺寸，取值为整数（单位为像素）。

⑥ 标记面填充颜色：'MarkerFaceColor'指定用于填充标记符面的颜色，颜色配比方案见表 1-2。例如，'MarkerFaceColor'，[0 1 0]表示标记面填绿色。

⑦ 标记周边颜色：如 ‘MarkerFaceColor’，'k'表示标记周边用黑色，其参数也见表 1-2。

表 1-2 LineSpec 可选字符串列表

线型			
标识符	意义	标识符	意义
-	实线	--	虚线
-.	点画线	:	点线
颜色			
标识符	意义	标识符	意义
r	红色	m	洋红色
g	绿色	y	黄色
b	蓝色	k	黑色
c	蓝绿色	w	白色
数据点标记类型			
标识符	意义	标识符	意义
+	加号	∧	向上的三角形
o	圆圈	∨	向下的三角形
*	星号	<	向左的三角形
.	点	>	向右的三角形
x	交叉符号	pentagram (或 p)	五边形
square (或 s)	方格	hexagram (或 h)	六边形
diamond (或 d)	菱形		

【例 1-17】 plot 函数示例 2。

```
t=0:pi/20:2*pi;
plot(t,t.*sin(t),'-.r*');
hold on;
plot(exp(t/100).*cos(t-pi/2),'--mo');
plot(sin(t*pi),':bs');
hold off
```

运行程序，效果如图 1-7 所示。

注意： hold on 表示继续在当前图形上画图。

【例 1-18】 plot 函数示例 3。

```
t=0:pi/20:2*pi;
plot(t,sin(2*t),'-mo','LineWidth',1.5,'MarkerEdgeColor','k','MarkerFaceColor',...
[0.49,1,0.63],'MarkerSize',10);
```

运行程序，效果如图 1-8 所示。

（2）fplot 函数

其调用格式如下：

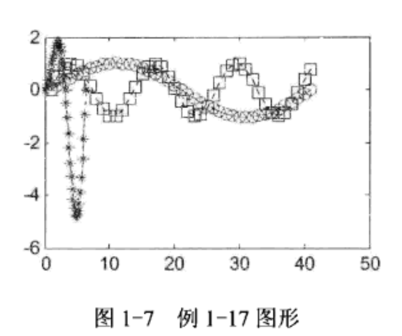


图 1-7 例 1-17 图形

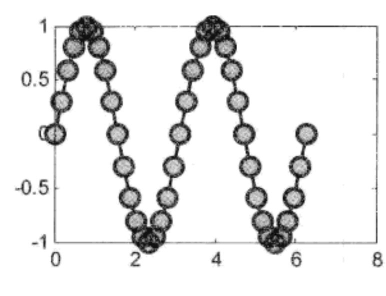


图 1-8 例 1-18 图形

```
fplot('function', limits)
```

其中，fplot('function', limits)：在指定的范围 limits 内画出函数名为 function 的一元函数图形。其中，limits 是一个指定 x 轴范围的向量[xmin, xmax]，或者是 x 轴和 y 轴的范围的向量[xmin, xmax, ymin, ymax]。

【例 1-19】 fplot 函数示例。

```
x=0:pi/18:2*pi;
fplot('sin(3*x)',[0 pi]);   %画出 x 在 0~pi 之间的 y=sin3x 的图形
fplot('[sin(x),cos(x)]',[-2*pi,2*pi]); %在同一张图上绘制正弦、余弦曲线
```

运行程序，效果如图 1-9 所示。

（3）符号函数 ezplot

其调用格式如下：

```
ezplot(f,[a, b])
```

其中，ezplot(f,[a, b])：绘出符号函数 f 在 a~b 区间的图形。

【例 1-20】 符号函数的绘图示例。

```
y=sym('cos(x)');
ezplot(y,[-2*pi,2*pi]);   %画出 x 在-2*pi,2*pi 之间的 y=cosx 的图形
```

运行程序，效果如图 1-10 所示。

（4）绘制对数图形函数 loglog、semilogx、semilogy

其调用格式如下：

```
loglog(X, Y)
semilogx(X, Y)
semilogy(X, Y)
```

其中，loglog(X, Y)：对 x 轴、y 轴的刻度用常用对数值（以 10 为底）；semilogx(X, Y)：对 x 轴的刻度用常用对数值，而 y 轴为线性刻度；semilogy(X, Y)：对 y 轴的刻度用常用对数值，而 x 轴为线性刻度。

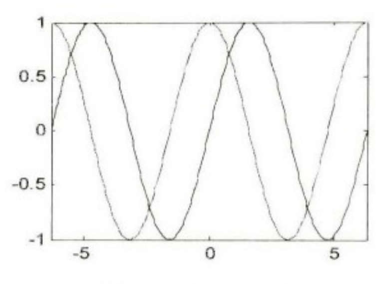


图 1-9　例 1-19 图形

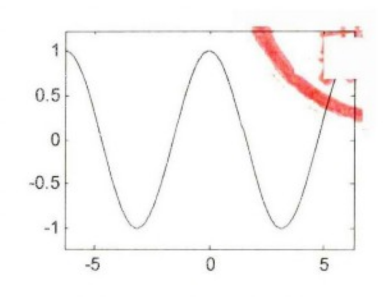


图 1-10　例 1-20 图形

【例 1-21】 绘制对数图形示例。

```
x=logspace(-1,2);
loglog(x,10*exp(x),'-s');
grid on;
```

运行程序，效果如图 1-11 所示。

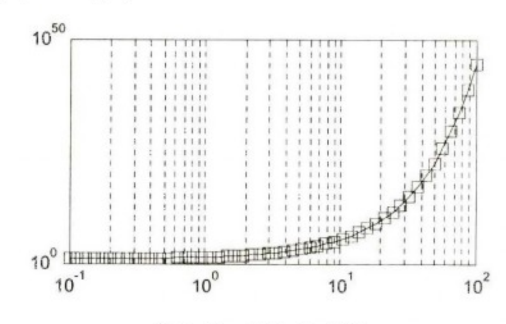


图 1-11　例 1-21 图形

（5）图形修饰与控制

```
axis square;          %将图形设置为正方形
axis equal;           %x,y 轴单位刻度相等
title('字符串');       %图形标题
axis([xmin,xmax,ymin,ymax]);     %x 轴范围为 xmin~xmax,y 轴范围为 ymin~ymax
xlabel('字符串');      %x 轴标注
ylabel('字符串');      %y 轴标注
text(x,y,'字符串');    %在(x,y)处标注说明文字
grid on;              %加网格线
grid off;             %消除网格线
hold on;              %保持当前图形
hold off;             %解除 hold on 命令
legend('First','Second',n);      %对一个坐标系上的两幅图形做出图例注解
subplot(m,n,p);       %将当前窗口分成 m 行 n 列区域，并指定在 p 区绘图
```

【例 1-22】 图形修饰与控制示例。

```
x=0:pi/60:2*pi;
```



```
subplot(221);plot(x,exp(-i*x));
subplot(222);fplot('log(x)',[10,2e3]);
subplot(212);plot(x,sin(x),':b',x,cos(x),'-r');
legend('sin(x)','cos(x)',1);
```

运行程序，效果如图 1-12 所示。

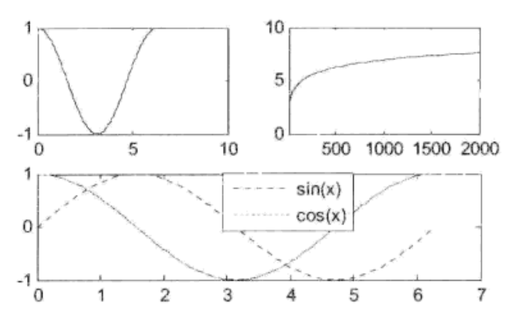


图 1-12　例 1-22 图形

注意：第二行语句 exp(-i*x)中的虚部被忽略；第三行语句中 2e3 表示 2×10^3，不能用 2*e3 表示，如 10^5 不能用 e5 表示，而用 1e5 表示；第四行语句 subplot(212)巧妙地将第二行整个区域用一个图形覆盖。

【例 1-23】 将正弦曲线 0~π/2 部分与轴包围的封闭图形填充为蓝色。

```
x=0:pi/60:2*pi;
y=sin(x);
x1=0:pi/60:pi/2;
y1=sin(x1);
plot(x,y,'-r');
hold on;
fill([x1,pi/2],[y1,0],'b');
```

运行程序，效果如图 1-13 所示。

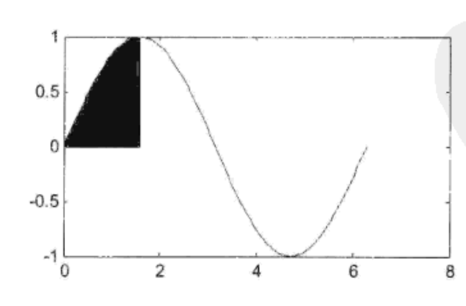


图 1-13　例 1-23 图形

（6）特殊二维图形

```
line([x1,x2],[y1,y2],'color',[0 0 1]);          %在(x1,y1)~(x2,y2)之间画一条蓝色直线
```

```
polar(theta,r);                    %用极角 theta 和极径 r 画出极坐标图形
pix(X);                            %绘制饼图
bar(X);                            %绘制条形图
stairs(X,Y);                       %绘制梯形图
```



2．绘制三维图形

（1）绘制三维曲线图函数 plot3、comet3、fill3

plot3 的调用格式如下：

```
plot3(X,Y,Z,S)
```

其中，plot3(X,Y,Z,S)：当 *X*，*Y*，*Z* 均为同维向量时，则 plot3 描出点 *X*(*i*)，*Y*(*i*)，*Z*(*i*)依次相连的空间曲线；若 *X*，*Y* 均为同维矩阵，*X*，*Y*，*Z* 每一组相应列向量为坐标画出一条曲线。*S* 为'color-linestyle-marker'控制字符。

【例 1-24】 绘制螺旋线。

```
t=0:pi/60:10*pi;
x=sin(t);
y=cos(t);
plot3(x,y,t,'*-r');
grid on;
```

运行程序，效果如图 1-14 所示。

comet3 的调用格式如下：

```
comet3(x, y, z)
```

其中，comet3(x, y, z)：显示一个彗星通过数据 *x*, *y*, *z* 确定的三维曲线。

【例 1-25】 函数 comet3 示例。

```
t=-20*pi:pi/50:20*pi;
comet3(sin(t),cos(t),t)
```

运行程序，效果如图 1-15 所示。

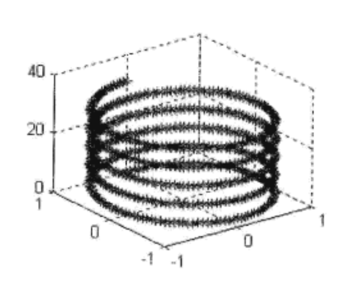


图 1-14　例 1-24 图形

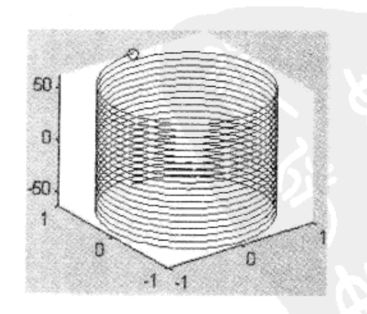


图 1-15　例 1-25 图形

可见到彗星头（一个圆圈）沿着数据指定的轨道前进的动画图像，彗星轨道为整个函数所画的螺旋线。

fill3 的调用格式如下：

```
fill3(X,Y,Z,C)
```

其中，fill3(X,Y,Z,C)：填充由参数 X，Y，Z 确定的多边形，参数 C 指定颜色。

【例 1-26】 fill3 函数示例。

```
clear all;
x=[2 1 2;9 7 1;6 7 0];
y=[1 7 0;4 7 9;0 4 3];
z=[1 8 6;7 9 6;1 6 1];
c=[1 0 0;0 1 0;0 0 1];
fill3(x,y,z,c);
grid on;
```

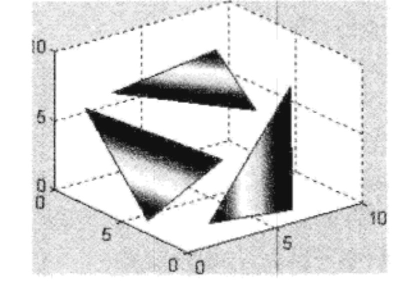


图 1-16 例 1-26 图形

运行程序，效果如图 1-16 所示。

（2）绘制三维网格图函数 mesh、meshc、meshz、meshgrid

mesh、meshc 和 meshgrid 的调用格式如下：

```
mesh(X,Y,Z,C)
meshc(X,Y,Z,C)
meshz(X,Y,Z,C)
```

其中，mesh(X,Y,Z,C)：画出颜色由 $\boldsymbol{C}$ 指定的三维网格图；meshc(X,Y,Z,C)：画出带有等高线的三维网格图；meshz(X,Y,Z,C)：画出带有底座的三维网格图。若 X 与 Y 均为向量，n=length(X)，m=length(Y)，Z 必须满足[m,n]=size(Z)，则空间中的点($X(j)$, $Y(i)$, $Z(i,j)$)为所画曲面网线的交点，X 对应于 Z 的列，Y 对应于 Z 的行。若 X，Y，Z 均为同维矩阵，则空间中的点($X(i,j)$, $Y(i,j)$, $Z(i,j)$)为所画曲面的网线的交点。矩阵 $\boldsymbol{C}$ 指定网线的颜色，MATLAB 对矩阵 $\boldsymbol{C}$ 中的数据进行线性处理，以便从当前色图中获得有用的颜色，若 $\boldsymbol{C}$ 缺省，网线颜色和曲面的高度 Z 相匹配。

在绘制三维图形时，常用到函数 meshgrid，其功能是生成二元函数 $z=f(x,y)$ 中 $x-y$ 平面上的矩形定义域中的数据点矩阵 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$。

[X,Y]= meshgrid(x,y)：输入向量 x 为 $x-y$ 平面上 x 轴的值，向量 y 为 $x-y$ 平面上 y 轴的值，输出矩阵 $\boldsymbol{X}$ 为 $x-y$ 平面上数据点的横坐标值，输出矩阵 $\boldsymbol{Y}$ 为 $x-y$ 平面上的数据点的纵坐标值。

【例 1-27】 meshgrid 函数示例。

```
>>x=1:4;
y=1:5;
[X,Y]=meshgrid(x,y)
X =
      1     2     3     4
      1     2     3     4
      1     2     3     4
```

```
      1     2     3     4
      1     2     3     4
Y =
      1     1     1     1
      2     2     2     2
      3     3     3     3
      4     4     4     4
      5     5     5     5
```

【例 1-28】 绘出带有底座的马鞍面。

$$Z=\frac{x^2}{4^2}-\frac{y^2}{5^2}$$

```
x=-8:8;
y=-8:8;
[X,Y]=meshgrid(x,y);
Z=(X.^2/4^2-Y.^2/5^2);
meshz(X,Y,Z);
```

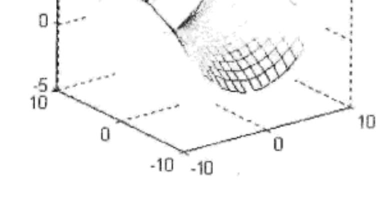


图 1-17　例 1-28 图形

运行程序，效果如图 1-17 所示。

（3）绘制三维曲面图 surf、surfc 函数

其调用格式如下：

```
surf(X,Y,Z,C)
surfc(X,Y,Z,C)
```

其中，surf(X,Y,Z,C)：画出颜色由 C 指定的三维曲面图；surfc(X,Y,Z,C)：画出带有等高线的三维曲面图。Surf 与 mesh 命令的用法和使用格式相同，不同之处在于，绘得的图形是一个彩色曲面而不是彩色网格。C 缺省时，数据 Z 为曲面高度，同时也是颜色数据。

【例 1-29】 绘出带有等高线的理想气体状态方程曲面（$v=nTR$，$n=2\text{mol}$）。

```
R=8.31;
n=2;
p=(1:20)*1e5;
v=(1:20)*1e-3;
[P,V]=meshgrid(p,v);
T=P.*V/n/R;
surfc(P,V,T);
view(45,45);
```

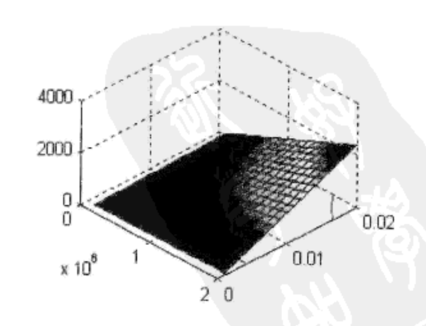


图 1-18　例 1-29 图形

运行程序，效果如图 1-18 所示。

（4）绘制三维旋转曲面图 cylinder 函数

其调用格式如下：

```
[X,Y,Z]=cylinder(r, n)
```

其中，[X,Y,Z]=cylinder(r, n)：返回母线向量为 r、高度为 1（见图 1-19 和图 1-20）的旋转曲面 x，y，z 轴的坐标值，旋转轴为 z 轴，旋转曲面的圆周有指定的 n 个距离相同的点。

用户可以用命令 surf 或命令 mesh 画出旋转的曲面图像。

【例 1-30】 绘制一个旋转抛物面 $z=\dfrac{(x^2+y^2)}{60}$。

```
z=0:20;
R=(60*z).^(1/2);
[X,Y,Z]=cylinder(R,30);
mesh(X,Y,Z);
```

运行程序，效果如图 1-19 所示。

（5）绘制三维球面图函数 sphere

其调用格式如下：

```
[X,Y,Z]=sphere(n)
```

其中，[X,Y,Z]=sphere(n)：生成三维直角坐标系中的单位球体坐标。该单位球体有 $n\times n$ 个面，该命令没有画图，只是返回矩阵，用户可以用命令 surf 或 mesh 画出球体。

【例 1-31】 绘制三维球面图示例。

```
z=0:25;
R=(60*z).^(1/2);
[X,Y,Z]=cylinder(R,30);
[X,Y,Z]=sphere;
mesh(X,Y,Z)
```

运行程序，效果如图 1-20 所示。

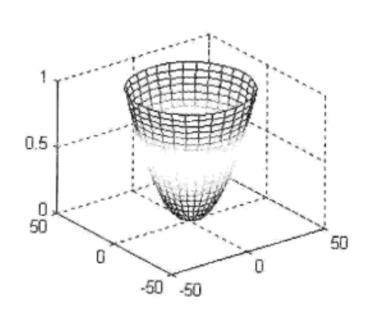


图 1-19　例 1-29 图形

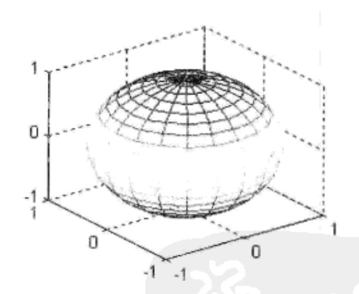


图 1-20　例 1-30 图形

1.1.6 MATLAB 数据类型及输出输入

1. 数据类型

MATLAB 的数据类型包括字符串、数值型（整型、单精度、双精度和稀疏矩阵）、单元数组、结构、Java 类和函数句柄。

（1）字符型（Char）和字符串（String）

字符型数组的元素是以 16 位无字符号整数表示的 Unicode ASCII 码。其中的 $1\times n$ 字符型数组，如 a='good'，又称为字符串。字符型数组除用符号的方法，也可以用 char 命令创

建，如 a=char('good')。

（2）数值型

数值型包括整型、单精度、双精度和稀疏矩阵等。其中，整型包括 8 位、16 位和 32 位的有符号和无符号整数型。在 MATLAB 中，整型要转化成双精度后才能进行数学运算。单/双精度浮点型意义与其他计算机高级语言一样。稀疏矩阵是一种特殊矩阵，它含有大量的零元素。MATLAB 只对矩阵中的非零元素进行存储和计算，可以用命令 sparse 创建稀疏矩阵。



（3）单元数组

单元数组是 MATLAB 数组的一种特殊数据类型。MATLAB 中，矩阵或数组只能保存相同类型或相同大小的数据，而单元数组允许把不同类型的 MATLAB 数组保存在不同的单元中。单元数组的每一个元素称为一个单元（Cell）。

单元数组有两种创建方式，一种是单元数组的各个元素直接赋值；另一种是先用 cell 函数为单元数组分配空间，然后再进行赋值。

【例 1-32】 单元数组示例。

```
a(1,1)={[2 2;1 1]};
a(1,2)={[0 0]};
a(2,1)={'cbcb'};
a(2,2)={3+j};
>> a
a =
      [2x2 double]              [1x2 double]
      'cbcb'              [3.0000 + 1.0000i]
>> a{1,1}
ans =
       2        2
       1        1
>> b=a{2,1}          %将单元数组值赋给 b
b = cbcb
>> bb=cell(2,2);     %先定义,然后再赋值
>> celldisp(a)       %显示单元数组的完整内容
a{1,1} =
       2        2
       1        1
a{2,1} = cbcb
a{1,2} =
       0        0
a{2,2} =
    3.0000 + 1.0000i
```

（4）结构

结构（Structure）也可以保存不同类型的数据，它由一组被称为域（Fields）的变量构成，数据存于域中。

结构的创建方法也有两种，一种是直接赋值；另一种是利用 struct 函数。例如：

```
>> group.name='A 组';
group.num='第五';
group.score=[70 83 92 100];
group.subject=['体能','身高','体重','营养'];
group                          %显示内容
group =
         name: 'A 组'
          num: '第五'
        score: [70 83 92 100]
      subject: '体能身高体重营养'
>> group(2).name='C 组';      %增加数组,在结构后面加下标
group(2).num='第六';
group(2).score=[75 82 91 99];
group(2).subject=['体能','身高','体重','营养'];
group(2)                        %显示新增内容
ans =
         name: 'C 组'
          num: '第六'
        score: [75 82 91 99]
      subject: '体能身高体重营养'
>> group
group =
1x2 struct array with fields:
      name
      num
      score
      subject
```

struct 命令的调用格式如下：

```
struct_array_name=struct('field1','values1', 'field2','values2',...)
```

其中，'field1'，'field2'，…代表域名（成员变量）；'values1'，'values2'，…代表对应的域值，其值必须是大小相同的单元数组、数量单元或单个数值。

（5）函数句柄

函数句柄也是一种数据类型。创建一个函数句柄，可用于保存函数的所有信息，以便将来对它进行调用。函数句柄可作为参数传递给其他函数，并与 feval 一起使用，以调用该函数句柄所属的函数。

创建函数句柄的调用格式如下：

```
funhandle=@function_name
```

其中，function_name 为用户指定的函数名；funhandle 为返回的函数句柄，可被另外的函数调用。

2. 数据输出

（1）fprintf 函数

此命令可按规定格式将数据输出到屏幕或写格式化数据到文件中。例如，fprintf('t%s\tx=%5d, \ty=%8.2f', x, y)函数，它包括两部分，即单引号括起来的部分及单引号后面的变量表。单引号内包含一些控制符，用于控制后面变量表中各变量的输出格式。例如，“\t”表示使光标移动一个制表位。“%”后跟“s”、“d”或“f”用于控制变量表中各变量的输出数据类型及所占的空格数。例如，“%”表示变量按字符串类型输出；“%5d”表示按整型类型输出，且共占 5 个空格；“%8.2f”表示按浮点类型输出，且共占 8 个空格。其他的非控制符按原样显示。



（2）disp 命令

此命令将结果输出到屏幕。

3. 数据输入

（1）利用 M 文件产生数据文件

利用文本编辑器可产生一个扩展名为*.m 文件，用于保存已知参数。调用此文件，就可将有关变量及其数据直接调入 MATLAB 内存中。

（2）fscanf 命令

此命令可从磁盘或文件中读取格式化数据，所用的控制符及用法与 fprintf 函数类似。

（3）input 命令

此命令提示用户从键盘中输入命令中提示的变量的值。

1.2 MATLAB 的程序编制

1.2.1 关系及逻辑运算

关系运算符主要用来比较数与数、矩阵与矩阵之间的大小，并返回真（用“1”表示）或假（用“0”表示）。基本的关系运算符主要有 6 种：>（大于）、<（小于）、≥（大于或等于）、≤（小于或等于）、==（等于）、~=（不等于）。

【例 1-33】 关系运算示例。

```
>> a=2<4          %小于运算
a =      1
>> b=4>=7         %大于或等于运算
b =      0
>> c=4==3         %等于运算
c =      0
>> d=4~=3         %不等于运算
d =      1
```

逻辑运算符有 4 种：与（&）、或（|）、非（~）、异或（xor）。在变量中，非零数的逻辑量为“真”，0 的逻辑量为“假”，逻辑运算结果以“1”表示“真”，以“0”表示“假”。

【例 1-34】 逻辑运算示例。

```
>> a=3&0          %“与”运算，两个真值间的结果为 1，否则为 0
a1=3|0            %“或”运算，有一个值为真或两个值都为真，则结果为 1
```

```
a2=xor(4,3)          % "异或"运算，只有一个值为真，则结果为 1，否则为 0
a3=~5                % "非"运算，真变假，假变真
a4=~0
```

运行结程序，输出如下：

```
a =     0
a1 =     1
a2 =     0
a3 =     0
a4 =     1
```

1.2.2 M 函数文件

sin(*x*)，sum(*A*)都是 MATLAB 内嵌的库函数，可以反复调用，十分方便。用户在实际工作中，往往需要编制自己的函数，以实现计算中的参数传递和函数的反复调用。建立函数文件的方法如下。

```
function[y1, y2,...]=ff(x1,x2,...)
```

其中，*ff* 是函数名；*x*1，*x*2 是输入变量；*y*1，*y*2 是输出变量。

【例 1-35】 计算一个向量所有元素的平均值。

```
%定义 aver.m 函数
function y=aver(x)
% 计算向量元素的平均值
% aver(x)为一个向量 x 元素的平均值
% 如果没有输入向量，程序将报错
[m,n]=size(x);
if(~((m==1)|(n==1))|(m==1 & n==1))
    error('please input a vector')
end
y=sum(x)/length(x)   %计算
```

这个例子包含了典型的 M 函数的各个部分：函数定义行、H1 行、帮助文档、函数主体和注释。

函数编辑完成后，将文件保存为 aver.m。程序中要求有一个输入参数，在命令行中输入 z 个向量并赋值：

```
>> z=1:199;
```

z=1:199 是定义函数的输入参数，输入文件名调用此函数：

```
>> aver(z)
```

运行程序，输出如下：

```
ans =    100
```



注意：① 输入变量用()括起来，输出变量用[]括起来。
② 函数名和文件名必须相同。函数名开头必须用字母，区分大小写。
③ 程序必须以 function 开始，第二行以后可加入注释行或运算语句。
④ M 函数文件可以调用其他一般的 M 文件，M 函数文件可以反复调用自己。
⑤ 用内联函数命令 inline 也可实现 M 函数文件的大部分功能。

【例 1-36】 内联函数示例。

```
>> fu=inline('2*x^2+2*x+1')   %默认 x 是输入参数
fv=inline('vo+a*t','a','t','vo');   %建立内联函数 fv,其中 vo,a,t 是变量
v=fv(4,5,2)   %求 a=4,t=5,vo=2 时函数 fv 的值
```

运行程序，输出如下：

```
fu =
       Inline function:
       fu(x) = 2*x^2+2*x+1
v =     22
```

1.2.3 M 文件

单击【File】菜单下的【New】子菜单中的【M-file】命令（或用桌面快捷键），进入文本编辑窗口，输入程序即可，开头可任意输入 MATLAB 语句。输完程序后，单击“保存”按钮，在对话框中输入文件名，文件名开头必须是字母。

运行 M 文件有以下几种方法：在命令窗口输入文件名并按〈Enter〉键；单击【File】菜单下的【Open】命令，在弹出的【Open】对话框中单击*.m（文件名），打开该文件编辑窗口，再单击【Debug】菜单下的【Run funl.m】命令即可。

1.2.4 程序控制语句

1. if 条件语句

格式一：

```
if（条件式）
     条件块语句组
end
```

例如，同循环语句举例相同的求和功能用条件转移语句实现，其程序如下。

```
msum=0;
for i=1:120
     if(msum>=5050)
          i
          msum
          break;
     end
```

```
        msum=msum+i;
    end
```

执行结果：

```
    i =
       101
    msum =
            5050
```

格式二：

```
    if（条件式）
        条件块语句组 1
    else
        条件块语句组 2
    end
```

格式三：

```
    if（条件式 1）
        条件块语句组 1
    elseif 条件式 2
        条件块语句组 2
    end
```

注意：① 在格式一中，表达式值非 0 时，执行下面语句；否则跳过，执行 end 后面的语句。

② 在格式二中，表达式值非 0 时，执行语句 1；否则执行语句 2。

③ 在格式三中，表达式值非 0 时，执行语句 1 并终止 if 语句；否则计算表达式 2 的值，依此类推。

【例 1-37】 比较数的大小。

```
a=3;b=6;
if a>b                  %条件表达式 1
    max=a;              %语句 1
elseif a==b             %条件表达式 2
    max='两数相等';      %语句 2
else
    max=b;              %语句 3
    disp(['最大值为:',num2str(max)]);
end
```

注意：if 和 end 必须成对使用；disp 的使用方法主要有 disp('…')和 disp(['…'])两种。

2. for 循环语句

格式：

```
for 循环变量=表达式 1：表达式 2：表达式 3
```

```
    循环语句组
end
```

注意：循环次数一般是给定的，除非用其他语句将循环提前结束（如 break）；表达式是一个向量；for 语句一定要有 end 作为结束标志；循环语句中的“;”可防止中间结果输出；循环体中，可以多次嵌套 for-end 结构体，但会影响运算速度。



【例 1-38】 利用 for-end 循环语句求出 100～200 之间的所有素数。

```
for m=101:2:200
    k=fix(sqrt(m));
    for i=2:k+1
        if rem(m,i)==0;
            break;
        end
    end
    if i>=k+1
        disp(int2str(m));
    end
end
```

3. while 循环语句

格式：

```
while（条件式）
    循环体条件组
end
```

注意：表达式一般由逻辑运算和关系运算组成。若表达式的值非 0，继续循环；若表达式的值为 0，中止循环。while 语句一定要有 end 作为结束标志。

【例 1-39】 用 while-end 循环语句求 1～100 之间整数的和。

```
>> sum=0;
i=1;
while i<=100
    sum=sum+i;
    i=i+1;
end
sum
```

运行程序，输出如下：

```
sum =        5050
```

4. switch 分支选择语句

这种语句是多分支选择语句，虽然有时可以用 if 语句的多层嵌套来完成，但没有 switch 语句显得简单明了。

格式：

```
switch  表达式
    case 常量表达式 1
        语句块 1
    case 常量表达式 2
        语句块 2
    …
    case 常量表达式 n
    otherwise
        语句块 n+1
end
```

注意：① switch 后面的表达式可以为任意类型。
② 当表达式的值与 case 后面的常量表达式的值相等时，就执行 case 后面的语句块。
③ case 后面的常量表达式可以有多个，也可以是不同类型。
④ 每次只执行一个语句块，执行完一个语句块就退出 switch 语句。

例如：

```
switch var
    case{'abc','12'}
        disp('第一种情况');
    case {1,2,4,'www'},
        disp('第二种情况');
    case {6,7,8,'MATLAB'},
        disp('第三种情况');
    otherwise
        disp('意外的情况');
end
```

注意：case 后面是{}，而不是()，运行结果为：
var=4，显示第二种情况。
var='abc'，显示第一种情况。
var=13，显示意外的情况。

1.2.5 编程要点

为了尽量加快 MATLAB 程序的运行速度，编程时应注意以下要点。

1）尽量避免使用循环，而使用向量或矩阵。

2）如果要使用循环，在循环语句前也要尽量对向量、矩阵或数组预先用 ones 或 zeros 函数进行内存分配。

3）尽量使用 MATLAB 的内部函数或工具箱函数。绝大多数常见的数学计算都可以在 MATLAB 中找到相应的函数命令。

在实际中，可以通过 tic（启动秒表）和 toc（停止秒表）测试程序运行所花费的时间。



第 2 章　概率与数理统计基本概念

概率论主要研究随机现象与其在数量方面的规律性，是数学的一个重要分支学科，现已广泛地应用于自然科学和社会人文科学的各个领域，成为处理信息、制定决策的重要理论基础。

数理统计是研究和解释随机现象统计规律性的一门数学学科。随机现象是指在个别试验中有可能发生，也有可能不发生，呈现不确定性，而在大量重复试验中又呈现统计规律性的一类现象。人们在科学实践活动中，经常接触到大量的随机现象，因此在科学研究中应用数量统计方法受到了人们的普遍重视。

2.1　随机事件及其概率

2.1.1 随机事件

自然界有许多现象，完全可以预言它们在一定条件下是否会出现。

例如，“同性电荷互相排斥”、“在一个标准大气压下，水加热到 100℃时必定沸腾”等是一定会出现的，而“同性电荷互相吸引”、“在一个标准大气压下，水加热到 100℃时不沸腾”等是必然不会出现的。

在一定条件下必然出现的现象称为必然事件，记为 Ω 。在一定条件下必然不出现的现象称为不可能事件，记为 $\varnothing$。显然，必然事件的反面就是不可能事件。

然而自然界还有许多现象，它们在一定的条件下可能出现，也可能不出现。

例如，“投掷一枚 1 元硬币（正面向上)”、“明天平均气温为 10℃”等就可能出现，也可能不出现。

粗略地讲，在一定条件下可能出现，也可能不出现的现象称为随机事件，或简称为事件，记为 A，B，C，…。

为了方便起见，后面将必然事件 Ω 和不可能事件 $\varnothing$ 也看做随机事件。

【例 2-1】 一次投掷两枚 1 元硬币，则：

$A=$“两枚都是正面朝上”。

$B=$“两枚都是正面朝下”。

$C=$“一枚正面朝上，一枚正面朝下”。

$D=$“至少有一枚正面朝下”。

都是随机事件。

【例 2-2】 设有 12 件产品，其中 9 件正品，3 件次品。现任意抽取 5 件，则：

$A=$“5 件都是正品”。

$B=$“至少有 1 件次品”。

C = “5 件都是次品”。

D = “至少有 1 件正品”。

都是随机事件，而 C 为不可能事件，D 则为必然事件。

2.1.2 概率

对于随机事件，在一次试验中是否发生，虽然不能预先知道，但是它们在一次试验中发生的可能性是有大小之分的。

比如，例 2-1 中的随机事件 A 和随机事件 B 发生的可能性是一样的，并且它们比随机事件 C 发生的可能性要小，既然各随机事件发生的可能性有大有小，自然使人想到该用一个数字 $P(A)$ 来标志随机事件 A 发生的可能性，较大的可能性用较大的数字来标志，较小的可能性就用较小的数字来标志。这个数字 $P(A)$ 就称为随机事件 A 的概率。

然而，对于已给的随机事件 A，到底应该用哪个数字来作为它的概率呢？也就是说，怎样从大小上来规定 $P(A)$ 呢？这决定于随机事件 A 的特殊性，不能一概而论。

对于随机事件 A，如果在一定条件下的 n 次试验中出现了 μ 次，则称 μ 为随机事件 A 在 n 次试验中出现的频数，并称此值：

$$f_n(A)=\frac{\mu}{n} \tag{2-1}$$

为随机事件 A 在 n 次试验中出现的频数。如果当试验次数 n 逐渐增大时，频率 $f_n(A)$ 在一个常数 p 附近摆动，而且逐渐稳定于这个常数 p，则称这种现象为频率的稳定性，而称常数 p 为频率稳定值。

例如，在一定条件下做投掷一枚 1 元硬币的试验，规定如下：

“硬币放在手心上，用一定的动作向上抛，使硬币自由地落在地面上……”。这些条件也称为条件组 S。于是，在条件组 S 的一次实现下，随机事件 A 是否发生是不确定的，然而这只是问题的一方面，当条件组 S 大量重复实现时，随机事件 A 发生的次数就能体现出一定的规律性，事实上约占总试验次数的一半。这可以写成

$$f_n(A)=\frac{频数}{试验次数}\approx\frac{1}{2} \tag{2-2}$$

即随机事件 A 具有频率的稳定性，且其频率稳定值为 $\frac{1}{2}$。

用 MATLAB 实现的投掷骰子实验。

用计算机模拟 100 次投掷一枚均匀骰子的实验结果，并写出相应的 MATLAB 命令代码。

在命令窗口中输入：

```
>> unidrnd(6,1,100)
```

输出如下：

```
ans =
  Columns 1 through 18
     5    6    1    6    4    1    2    4    6    6    1    6    6    3    5    1    3    6
  Columns 19 through 36
```

```
  5   6   4   1   6   6   5   5   5   3   4   2   5   1   2   1   1   5
Columns 37 through 54
  5   2   6   1   3   3   5   5   2   3   3   4   5   5   2   5   4   1
Columns 55 through 72
  1   3   6   3   4   2   5   2   4   5   6   6   4   1   1   2   6   2
Columns 73 through 90
  5   2   6   3   2   2   4   3   3   5   4   4   6   2   5   5   3   4
Columns 91 through 100
  1   1   4   5   6   1   4   3   1   3
```



模拟结果介于 1～6 之间，与掷出骰子实验结果相对应。

而在历史上，有些人曾做过成千上万次投掷硬币的试验。表 2-1 列出了其试验记录。

表 2-1 硬币试验记录

试 验 者	投掷次数 n	出现“反面朝上”次数 μ	频率 $\frac{\mu}{n}$
Pearson	24000	12012	0.5005
Pear	12000	6019	0.5016
Buffon	4040	2048	0.5069
Demorgan	2048	1061	0.518

从表 2-1 中容易看出，投掷次数越多，频率越接近 0.5。

人类的大量实践证明，在实际中所遇到的随机事件，一般都具有频率的稳定性，因此，所谓某事件发生的可能性的大小，在数量上可以用“频率稳定值”来刻画。

定义 2-1 在一组不变的条件 S 下，随机事件 A 的频率稳定值 p 就称为随机事件 A 在条件组 S 下发生的概率，记为 $P(A)$，即

$$P(A)=p \tag{2-3}$$

由于频率 $f_n(A)$ 总介于 0～1 之间，因而由概率的定义 2-1 知，对任何随机事件 A 有

$$0 \leqslant P(A) \leqslant 1 \tag{2-4}$$

而对必然事件 Ω 及不可能事件 $\varnothing$，则显然有

$$P(\Omega)=1, \quad P(\varnothing)=0 \tag{2-5}$$

【例 2-3】 计算机模拟 1000 次投掷一枚均匀骰子的实验结果。对于 $i=1,2,\cdots,10$，以及 $i=50100$，分别写出前 $i\times10$ 次各个结果出现的频率，观察频率随实验次数增加的变化规律，并写出完成上述任务的 MATLAB 命令代码。

在 MATLAB 命令行中，输入以下代码：

```
>> x=unidrnd(6,1000,1);y=x(1:20)
f1=sum([y==1,y==2,y==3,y==4,y==5,y==6])/10;
y=x(1:20)
f2=sum([y==1,y==2,y==3,y==4,y==5,y==6])/20;
y=x(1:30)
f3=sum([y==1,y==2,y==3,y==4,y==5,y==6])/30;
y=x(1:40)
```

```
f4=sum([y==1,y==2,y==3,y==4,y==5,y==6])/40;
y=x(1:50)
f5=sum([y==1,y==2,y==3,y==4,y==5,y==6])/50;
y=x(1:60)
f6=sum([y==1,y==2,y==3,y==4,y==5,y==6])/60;
y=x(1:70)
f7=sum([y==1,y==2,y==3,y==4,y==5,y==6])/70;
y=x(1:80)
f8=sum([y==1,y==2,y==3,y==4,y==5,y==6])/80;
y=x(1:90)
f9=sum([y==1,y==2,y==3,y==4,y==5,y==6])/90;
y=x(1:100)
f10=sum([y==1,y==2,y==3,y==4,y==5,y==6])/100;
y=x(1:500)
f50=sum([y==1,y==2,y==3,y==4,y==5,y==6])/500;
y=x
f100=sum([y==1,y==2,y==3,y==4,y==5,y==6])/1000;
```

表 2-2 列出了最终的频率计算结果。从该表中可以看出：随着实验次数的增加，掷出的各个点数的频率接近于 $1/6 \approx 0.16667$，即频率随着实验次数的增加而稳定于概率值。

表 2-2　模拟 1000 次投掷均匀骰子实验的频率统计结果

骰子点数 / 频率 / 模拟次数	1	2	3	4	5	6
10	0.00000	0.00000	0.30000	0.20000	0.10000	0.40000
20	0.10000	0.10000	0.20000	0.15000	0.10000	0.35000
30	0.13333	0.10000	0.16667	0.16667	0.10000	0.33333
40	0.15000	0.07500	0.17500	0.20000	0.12500	0.27500
50	0.14000	0.10000	0.20000	0.18000	0.14000	0.24000
60	0.15000	0.11667	0.16667	0.18333	0.15000	0.23333
70	0.15714	0.11429	0.14286	0.15714	0.18571	0.24286
80	0.15000	0.10000	0.12500	0.16200	0.23750	0.22500
90	0.13333	0.13333	0.13333	0.16667	0.23333	0.20000
100	0.13000	0.12000	0.14000	0.16000	0.23000	0.22000
500	0.16000	0.14400	0.16400	0.14800	0.20200	0.18200
1000	0.16600	0.15000	0.16400	0.16200	0.18400	0.17400

一般地，对于一个行（列）向量 $\boldsymbol{x}$，给定正整数 i，代码 x(i)表示 x 的第 i 个分量；而给定一个以正整数为分量的 n 维向量 $\boldsymbol{i}$，代码 x(i)表示 n 维行（列）向量，其第 k 个分量为 x(i(k))。当然，这里要求 $\boldsymbol{i}$ 的各个分量都不超过 $\boldsymbol{x}$ 的维数。例如，代码

```
>> x=1:6;
i=[1,1,2,2,3,3,4,4,5];
x(i)
```

输出如下：

```
ans =
     1     1     2     2     3     3     4     4     5
```



即得到一个9维的行向量（1, 1, 2, 2, 3, 3, 4, 4, 5）。在上面的这段代码中，$\boldsymbol{x}$是一个5维的向量，而$\boldsymbol{i}$是一个9维的向量，它的各个分量的取值都是小于或等于5的正整数，因此代码x(i)表示一个9维的行向量。由于$\boldsymbol{i}$的第一个分量是1，所以x(i)的第一个分量等于$\boldsymbol{x}$的第一个分量1；由于$\boldsymbol{i}$的第二个分量也是1，所以x(i)的第二个分量等于$\boldsymbol{x}$的第一个分量1；由于$\boldsymbol{i}$的第三个分量是2，所以x(i)的第三个分量等于$\boldsymbol{x}$的第二个分量2；由于$\boldsymbol{i}$的第四个分量还是2，所以x(i)的第四个分量等于$\boldsymbol{x}$的第两个分量2；……；由于$\boldsymbol{i}$的第九个分量是5，所以x(i)的第九个分量等于$\boldsymbol{x}$的第五个分量5。

当x是一个$m \times n$阶矩阵时，也可以用类似的代码表示这个矩阵的某个位置的元素，或由这个矩阵的某些行或列的交叉位置的元素组成的矩阵。例如，代码：

```
>> a=[1 2 3;4 5 6;7 8 9];
a(1,2)
```

输出如下：

```
ans =     2
```

即得到矩阵$\boldsymbol{a}$的第一行和第二列交叉位置的元素2。如果继续输入代码

```
>> i=[1 3];
j=[1 2];

a(i,j)
```

输出如下：

```
ans =
     1     2
     7     8
```

这个结果恰好是由矩阵$\boldsymbol{a}$的第一、三行与第一、二列交叉位置的元素所构成的2×2阶矩阵。也就是说，i指定了$\boldsymbol{a}$中参与构成矩阵$\boldsymbol{a}(i, j)$的行；而j指定了$\boldsymbol{a}$中参与构成矩阵$\boldsymbol{a}(i, j)$的列。

2.1.3 排列与组合

为了计算随机事件的概率，下面介绍排列与组合的基本知识，首先给出两个基本原理。

1. 两个基本原理

（1）加法原理

完成某项工作，有两类不同的方法：方法甲与方法乙。方法甲有m种方式，方法乙有n种方式，都可以完成这项工作，则完成该项工作有$m+n$种方式。

（2）乘法原理

完成某项工作，必须通过两个步骤。第一个步骤有 m 种方式，第二个步骤有 n 种方式，则完成该项工作共有 mn 种方式。

2．排列

（1）选排列

从 n 个不同元素中任选 r 个元素（不允许重复，$r<n$）按照一定顺序排成一列，称为从 n 个不同元素中取 r 个元素的一个选排列。其排列总数用 A_n^r 来表示，则有

$$A_n^r = n(n-1)\cdots(n-r+1) = \frac{n!}{(n-r)!} \tag{2-6}$$

（2）全排列

将 n 个不同元素按照一定顺序排成一列，称为这 n 个不同元素的一个全排列。其排列总数用 P_n 来表示，则有

$$P_n = n(n-1)\cdots 3\cdot 2\cdot 1 = n! \tag{2-7}$$

（3）可重复的排列

从 n 个不同元素中任取 r 个元素（允许重复，$r\leqslant n$）按照一定顺序排列成一列，称为一个可重复的排列。其排列总数用 U_n^r 来表示，则有

$$U_n^r = \underbrace{n\cdot n\cdots n}_{r个n} = n^r \tag{2-8}$$

3．组合

从 n 个不同元素中任取 r 个元素（不允许重复，$r\leqslant n$）不计顺序构成一组，称为从 n 个不同元素中取 r 个元素构成的一个组合。其组合总数用 C_n^r 来表示，则有

$$C_n^r = \frac{A_n^r}{r!} = \frac{n(n-1)\cdots(n-r+1)}{r!} = \frac{n!}{r!(n-r)!} \tag{2-9}$$

4．排列与组合的示例

【例 2-4】 从 26 个英文字母中取 3 个不同的字母组成单词，最多能组成多少个单词？

解： $A_{26}^3 = 26\times 25\times 24$ 个 $=15600$ 个

【例 2-5】 由数字 0, 1, 2, 3, 4, 5 能组成多少个没有重复数字的 5 位数？

解： 因为首位数不能为 0，所以该位上只有 5 种选择，其余 4 位数由剩下的 5 个数字进行排列，共有 A_5^4 种选择，所以组成没有重复数字的 5 位数共有

$$5A_5^4 = 5\times 5\times 4\times 3\times 2 \text{ 个} = 600 \text{ 个}$$

【例 2-6】 从 0, 1, 2,…,9 共 10 个数字中取出 8 个数字组成电话号码，求

1）8 个数字均不相同的电话号码总数。

2）8 个数字组成的所有可能的电话号码总数。

解： 1）利用选排列的计算公式得

$$A_{10}^8 = \frac{10!}{2!} \text{ 个} = 1814400 \text{ 个}$$

2）利用可重复排列的计算公式得

$$U_{10}^{8}=10^{8}个$$

【例 2-7】 平面上有 10 个点，任何 3 点不共线，问

1）共能做成多少个三角形？

2）以其中一点 A 为顶点的三角形共有多少个？



解： 1）因为任何 3 点不共线，所以在 10 个点中任取 4 点都能做成三角形，共能做成

$$C_{10}^{3}=\frac{10!}{3!(10-3)!}个=120个$$

2）由于三角形的一个顶点 A 已定，所以其余两点只能在 9 个点中任取两点，于是所求的三角形总数为

$$C_{9}^{2}=\frac{9!}{2!(9-2)!}个=36个$$

2.1.4 古典概率

上面给出了概率的定义，它既是概念，同时又提供了近似计算概率的一般方法。在某些特殊情况下，并不需要临时做多次试验来求得概率的近似值，而是根据问题本身所具有的某种“对称性”，充分利用人类长期积累的关于“对称性”的实际经验，分析事件的本质，就可以直接计算其概率。

例如，投掷一枚 1 元硬币，即使不临时做大量的投掷试验，也会想到，“正面朝上”与“正面朝下”出现的机会相等。因此，可以推测在大量试验中“正面朝上”发生的频率在 1/2 左右，即其概率为 1/2。为什么“正面朝上”与“正面朝下”的机会相等呢？这是因为问题本身具有一定的对称性。如果“朝上”与“朝下”出现的机会不相等，那反倒与人类长期形成的“对称”经验不相符了。

【例 2-8】 盒中装有 3 个白球、2 个黑球共 5 个球，从中任取一个，问取到白球的概率是多少？

解： 既然是任取，那么取到每一个球的机会是一样的，而白球有 3 个，因此取到白球的概率是 3/5。

讲得更清楚些，可以把 5 个球编上号，其中 1、2、3 号为白球，4、5 号为黑球。因此任取一个，所以“取到 i 号球”（$i=1,2,3,4,5$）这 5 个结果发生的机会一样，而且是互相排斥的，除此之外不可能有别的结果。注意到 1、2、3 号球是白球，所以“取到白球”这一事件发生的频率会稳定在 3/5 左右，因此按照概率的定义，其概率是 3/5。

【例 2-9】 盒中装有球的情况同例 2-8，现从中任取 2 个，问 2 个球全是白球的概率是多少？

解： 这个问题较例 2-8 复杂，不过仍可按例 2-8 的方法进行分析，把 5 个球同样编号。因为是任取 2 球，所以“①②”，“①③”，“①④”，“①⑤”，“②③”，“②④”，“②⑤”，“③④”，“③⑤”，“④⑤”发生的机会一样，且互相排斥，除此之外不可能有别的结果。再注意到上面 10 种情况中，有且仅有 3 种，即

“①②”，“①③”，“②③”

为全白，因此“全白”发生的概率会稳定在 3/10 左右，即其概率是 3/10。

定义 2-2 如果一个事件组 $A_1, A_2, \cdots, A_n$ 具有下列 3 条性质：

1）等可能性：$A_1, A_2, \cdots, A_n$ 发生的机会相同。

2）完全性：在任一次试验中，$A_1, A_2, \cdots, A_n$ 中至少有一个发生。

3）互不相容性：在任一次试验中，$A_1, A_2, \cdots, A_n$ 中至多有一个发生。

则称事件组 $A_1, A_2, \cdots, A_n$ 为等概基本事件组。若称为等可能完备事件组，其中任一事件 $A_i (i=1,2,\cdots,n)$ 称为基本事件。

如果 $A_1, A_2, \cdots, A_n$ 是一个等概基本事件组，而事件 B 由其中的某 m 个基本事件 $A_{i_1}, A_{i_2}, \cdots, A_{i_m}$（$m \leqslant n$）所构成，则事件 B 的概率由下列式子来计算：

$$P(B) = \frac{m}{n} \tag{2-10}$$

利用式（2-10）来讲解等概基本事件组概率的模型，称为古典概型。

现通过式（2-10）再来讲解例 2-8。

从 3 个白球、2 个黑球中任取 2 个球，共有 $\mathrm{C}_5^2=10$ 种不同的取法，每一种取法对应一个事件，可以验证由这 10 种不同取法构成的事件组是一个等概基本事件组（验证这里从略），而取得 2 个球均为白球这一事件是由 $\mathrm{C}_3^2=3$ 种取法对应的 3 个基本事件所构成的，所以利用式（2-10）即得

$$P(\text{取得两个白球}) = \frac{\mathrm{C}_3^2}{\mathrm{C}_5^2} = \frac{3}{10}$$

下面再介绍两个古典概型的例子。

【例 2-10】 设有 m 件产品，其中有 k 件次品（$k \geqslant 2$，$m \geqslant 50+k$）。现从中任取 50 件，求下列事件的概率：

$A=$“无次品”；$B=$“恰有 2 件次品”。

解：从 m 件产品中任取 50 件，其有 C_m^{50} 种不同的取法，每一种取法对应一个事件，容易验证这些事件构成一个等概基本事件组（验证这里从略）。

显然，所要取的 50 件产品中无次品，必须是从 $m-k$ 件正品中取来的，可见这种无次品的取法共有 C_{m-k}^{50} 种，即事件 A 含有 C_{m-k}^{50} 个基本事件，所以由式（2-10）得

$$P(A) = \frac{\mathrm{C}_{m-k}^{50}}{\mathrm{C}_m^{50}}$$

取出的 50 件产品中，恰有 2 件产品，即有 48 件正品，2 件次品。这 48 件正品必是从 $m-k$ 件正品中取出的，共有 C_{m-k}^{48} 种取法；而 2 件次品必是从 k 件次品中取出的，共有 C_k^2 种取法。因此，事件 B 共包含 C_{m-k}^{48} C_k^2 个基本事件，于是根据式（2-10）得

$$P(B) = \frac{\mathrm{C}_{m-k}^{48}\mathrm{C}_k^2}{\mathrm{C}_m^{50}}$$

【例 2-11】 有 n 个人，每人以同样的概率 $\frac{1}{N}$ 被分配在 N（$n \leqslant N$）间房中的任一间中，求下列各事件的概率：

$A=$“某指定 n 间房中各有一人”；

$B=$“恰有 n 间房，其中各有一人”;

$C=$“某一指定间房中恰有 m（$m\leqslant n$）人”;

$D=$“恰有一间房中有 m（$m\leqslant n$）人”。



解：每个人都可以分配到 N 间房中的任一间中，共有 N 种不同的分法。n 个人分配到 N 间房中就有 $U_N^n=N^n$ 种不同的分法，即等概基本事件组含有 N^n 个基本事件。

现指定 n 间房，n 个人被分配到这 n 间房中去，每间房 1 人，共有 $P_n=n!$种分法，即事件 A 含有$n!$个基本事件，于是

$$P(A)=\frac{n!}{N^n}$$

如果这 n 间房可由 N 间房中任意选出，则共有 C_N^n 种选法，因而事件 B 共含有 $n!\,C_N^n$ 个不同的基本事件，于是

$$P(B)=\frac{n!C_N^n}{N^n}=\frac{N!}{N^n(N-n)!}$$

事件 C 中的 m 个人可从 n 个人中任意选出，共有 C_n^m 种选法，其余 $n-m$ 个人可以任意分配在其余的 $N-1$ 间房中，共有 $(N-1)^{n-m}$ 种分法，因而事件 C 共包含有 $C_n^m(N-1)^{n-m}$ 个不同的基本事件，于是

$$P(C)=\frac{C_n^m(N-1)^{n-m}}{N^n}=\frac{n!}{m!(n-m)!}\left(\frac{1}{N}\right)^m\left(1-\frac{1}{N}\right)^{n-m}$$

如果从 N 间房中任意选出一间，则有 N 种选法，因而事件 D 共包含有 $NC_n^m(N-1)^{n-m}$ 个不同的基本事件，于是

$$P(D)=\frac{NC_n^m(N-1)^{n-m}}{N^n}=\frac{n!}{m!(n-m)!}\left(\frac{1}{N}\right)^{m-1}\left(1-\frac{1}{N}\right)^{n-m}$$

【例 2-12】 考察某网站在 1h 内被点击次数的变化情况，用非负整数 n 表示结果“该网站在 1h 内被点击 n 次”，则样本点可以用非负整数表示，样本空间为

$$\Omega=\{0,1,2,\cdots\}$$

2.2 事件及运算

在随机现象的研究中，样本点是最小的研究单位，但用户对具有某种特性的样本点会不会很感兴趣。例如，在掷骰子实验中，“掷出的点数小于 3”也是一个可能出现的结果。

定义 2-3 部分样本点组成的结果称为随机事件，简称为事件，常用大写字母 $A,B,C,\cdots$ 表示；一定要发生的事件称为必然事件，用 Ω 表示；一定不发生的事件称为不可能事件，用 ϕ 表示。

注意：可以用集合的观点来看待事件，即事件是样本空行的一个子集。这个观点下，随后要介绍的事件之间的关系和事件之间的运算可以看成是集合之间的关系和集合之间的运算。

如在例 2-12 中，该网站“至多被点击 2 次”是一个事件。该事件由“没有被点击”、“被点击 1 次”和“被点击 2 次”3 个样本点所组成。在例 2-2 中，“5 件都是次品”为一个事件，它是不可能事件；“至少有 1 件正品”也为一个事件，它由所有的样本点组成，是必然事件。

从定义来看，事件是一种特殊的集合，可以用集合的方式来表达事件。如在例 2-12 中，该网站“至多被点击 2 次”的事件 A 可以用集合的形式表示为：

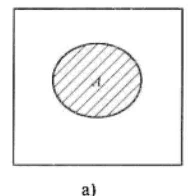
$A=\{$“没有被点击”，“被点击 1 次”，“被点击 2 次”$\}$

而“被点击 1 次”的事件 B 可以表示为：

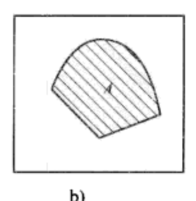
$B=\{$“被点击 1 次”$\}$

进一步，如果用数字 i 表示事件“被点击 i 次”，则上面的两个事件 A 和 B 可以分别表示为 $A=\{0,1,2\}$ 和 $B=\{1\}$，必然事件可以表示为 $\Omega=\{0,1,2,\cdots\}$。

可以用平面图形来图示事件，具体方法是用一个封闭的平面曲线的内部表示一个事件。例如，对于事件 A，图 2-1a 中矩形的内部表示样本空间（必然事件），椭圆型区域的内部表示事件 A，这样的示意图称为维恩图。注意，在维恩图中，表示事件的区域边缘（即封闭曲线的形状）可以是任意的，但它必须位于表示必然事件的区域内部，如图 2-1b 中的封闭曲线的内部区域也可以表示事件 A。



a)



b)

图 2-1　事件的示意图——维恩图

a) 事件 A 的维恩图　b) 事件 A 的维恩图

事件 A 发生或出现事件 A 是指出现了构成该事件的样本点 ω，即出现的样本点 $\omega\in A$。事件 A 不发生或没有出现事件 A 是指出现的样本点 ω 不在事件 A 中，即出现的样本点 $\omega\notin A$。

定义 2-4　如果事件 A 发生能够推出事件 B 发生，则称事件 A 包含于事件 B，或事件 A 被 B 包含，简称为 A 包含于 B，或 A 被 B 包含，记为 $A\subset B$；也可以称为事件 B 包含事件 A，简称为 B 包含 A，记为 $B\supset A$。如果 A 包含 B，且 B 包含 A，则称事件 A 等于事件 B，记为 $A=B$。

注意： 事件的包含关系、被包含关系和相等关系的定义与集合论中的相应概念类似。这里用事件发生或不发生来定义这些关系能更好地体现事件之间关系的实际含义。

定义 2-5　如是事件 A 和事件 B 没有公共的样本点，则称事件 A 和事件 B 互斥，或事件 A 和事件 B 不相容。如果 n 个事件（一个事件列）的任意两个事件都互斥，则称这 n 个事件（这个事件列）两两相斥，或两两不相容。

【例 2-13】 在例 2-12 中，用 A 表示事件“该网站在 1h 内被点击奇数次”，B 表示事件“该网站在 1h 内被点击 3 次”，C 表示事件“该网站在 1h 内被点击次数大于 5 次”，则

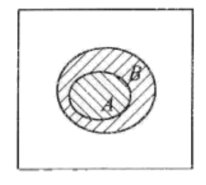
$$A=\{2n+1\,|\,n=0,1,2,\cdots\},\quad B=\{3\},\quad C=\{6,7,8,\cdots\}$$

并且 $B\subset A$，B 和 C 互斥，A 和 C 不互斥。

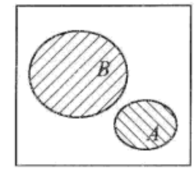
可以借助维恩图来理解和记忆事件的包含和互斥关系的含义。如在图 2-2a 中，代表事件 A 的区域完全落在代表事件 B 的区域内部，这表示 A 的样本点全部是 B 的样本点，即关系 $A\subset B$ 成立。在图 2-2b 中，代表事件 A 和事件 B 的区域没有公共部分，表示两个事件没

有公共的样本点，即它们互斥。

在实际应用中，人们经常用一些简单的事件来构造新的事件，这涉及事件的运算。下面介绍常用的几种事件的运算。



a)



b)

图 2-2　表示事件关系的维恩图

a) $A\subset B$ 或 $B\supset A$　b) A 与 B 互斥

（1）并

把事件 A 和 B 的所有样本点合到一起所构成的事件称为事件 A 与 B 的并，记为 $A\cup B$。可以把事件 A 与 B 的并表示为

$$A\cup B=\{\omega\mid\omega\in A\text{或}\omega\in B\}$$

显然，事件 $A\cup B$ 发生等价于事件 A 和事件 B 中至少有一个事件发生。

（2）交

用事件 A 和 B 所共有的样本点构成的事件称为事件 A 与 B 的交，记为 $A\cap B$，或 AB。可以把事件 A 与 B 的交表示为

$$A\cap B=\{\omega\mid\omega\in A\text{且}\omega\in B\}$$

显然，事件 $A\cap B$ 发生等价于事件 A 和事件 B 同时发生。

（3）差

在事件 A 中而不在 B 中的样本点所构成的事件称为事件 A 与 B 的差，记为 $A-B$。可以把事件 A 与 B 的差表示为

$$A-B=\{\omega\mid\omega\in A\text{且}\omega\notin B\}$$

显然，事件 $A-B$ 发生等价于在 A 和 B 两个事件中，仅有事件 A 发生。

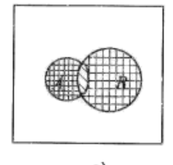
（4）补

由不在 A 中的样本点所构成的事件称为事件 A 的余事件，或事件 A 的补事件，记为 $\overline{A}$。可以把事件 A 的余事件（补事件）表示为

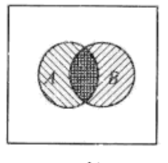
$$\overline{A}=\{\omega\mid\omega\notin A\}=\Omega-A$$

显然，事件 $\overline{A}$ 发生等价于事件 A 不发生；$\overline{A}$ 的余事件等于事件 A，即 $\overline{(\overline{A})}=A$。

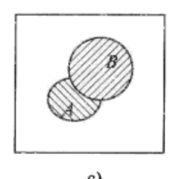
可以借助于维恩图来理解事件的运算，如图 2-3 所示。图 2-3a 中的网线区域表示事件 $A\cup B$，其含义为该事件由事件 A 和 B 的所有样本点构成；图 2-3b 中的网线区域表示事件 $A\cap B$，其含义为该事件由事件 A 和 B 的所有公共样本点构成；图 2-3c 中的斜线（即从左上到右下的斜线）区域表示事件 $A-B$，其含义为在事件 A 中去掉事件 B 中的所有样本点所剩余的那些样本点所构成的事件；图 2-3d 中的网线区域表示事件 A 的余事件，其含义为该事件由不在事件 A 中的样本点所构成。



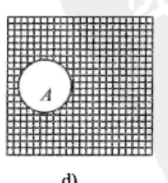
a)



b)



c)



d)

图 2-3　表示事件之间运算的维恩图

a) $A\cup B$　b) $A\cap B$　c) $A-B$　d) $\overline{A}$

【例 2-14】 在掷骰子实验中，用数字$i(i=1,2,3,4,5,6)$表示掷出的结果为“i 点面向上”，则样本空间为

$$\Omega=\{1,2,3,4,5,6\}$$

记$A=\{1,2,3\}$，$B=\{4,5,6\}$，$C=\{3,4,5,6\}$。

$$A\cup B=\Omega，\ AB=\varnothing，\ A-C=\{1,2\}$$

$$\overline{A}=B，\ \overline{B}=A，\ A-B=A$$

因此$A\cup B$为必然事件，AB 为不可能事件，$A-C$表示“出现的点数小于 3”，$\overline{A}$为事件B，$\overline{B}$为事件A，$A-B$还等于事件A本身。

定理 2-1 事件运算的对偶律。

1）$\overline{A\cup B}=\overline{A}\cap\overline{B}$。

2）$\overline{A\cap B}=\overline{A}\cup\overline{B}$。

证明： $\omega\in\overline{A\cup B}$ 等价于 $\omega\notin A\cup B$，即 $\omega\notin A$ 且 $\omega\notin B$，亦即 $\omega\in\overline{A}\cap\overline{B}$，所以$\overline{A\cup B}=\overline{A}\cap\overline{B}$，即结论 1）成立。

注意到$\overline{(\overline{A})}=A$，由结论 1）可得$\overline{\overline{A}\cup\overline{B}}=A\cap B$，等式两边再次取余运算可得结论 2)。

注意： 借助维恩图，容易理解事件运算的对偶律。事实上，事件$\overline{A\cup B}$的维恩图如图 2-4a 所示。其中的网线区域代表该事件，这个网状区域当然等于图 2-4b 中的斜线区域（代表事件$\overline{A}$）和图 2-4c 中的反斜线（即从右上到左下的斜线）区域（代表事件$\overline{B}$）的公共部分，即$\overline{A\cup B}=\overline{A}\cap\overline{B}$。类似地，图 2-4d 中的斜线区域代表事件$\overline{A\cap B}$，这个斜线区域是由图 2-4e 中的斜线区域（代表事件$\overline{A}$）和图 2-4f 中的斜线区域（代表事件$\overline{B}$）合并所构成的，即$\overline{A\cap B}=\overline{A}\cup\overline{B}$。

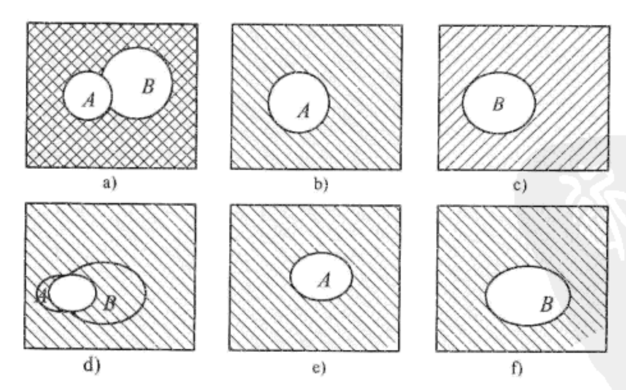


图 2-4 表示事件关系的维恩图

a) $\overline{A\cup B}$ b) $\overline{A}$ c) $\overline{B}$ d) $\overline{A\cap B}$ e) $\overline{A}$ f) $\overline{B}$

定理 2-2 事件运算的简单性质。

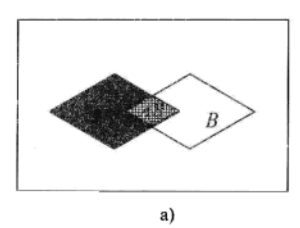
1）$A-B=A-(AB)=A\overline{B}$。

2）$(A\cup B)C=(AC)\cup(BC)$。

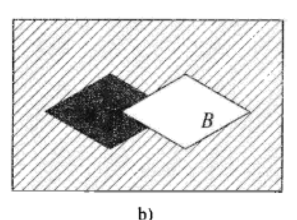
3）$(AB)\cup C=(A\cup C)(B\cup C)$。

证明：如图 2-5a 所示，AB 由事件 A 和事件 B 的公共样本点所构成（图中的网线区域），因而从事件 A 去掉事件 B 中的所有样本点等价于从 A 中去掉事件 AB 中的所有样本点，即 $A-B=A-(AB)$。如图 2-5b 所示，$\overline{B}$ 由不在事件 B 中的样本点所构成（图中的斜线区域），这些样点与事件 A 中样本点的公共部分恰好构成 $A-B$，因此 $A-B=A\overline{B}$。所以结论 1）成立。





a)



b)

图 2-5　$A-B=A-(AB)=A\overline{B}$ 的示意图

a) $A-B=A-(AB)$ 的示意图　b) $A-B=A\overline{B}$ 的示意图

如图 2-6 所示，由于 $A\subset A\cup B$，$B\subset A\cup B$，所以 $AC\subset(A\cup B)C$，$BC\subset(A\cup B)C$，进而有

$$(A\cup B)C\supset(AC)\cup(BC) \tag{2-11}$$

另一方面，若 $\omega\in(A\cup B)C$，则 $\omega\in A\cup B$ 且 $\omega\in C$。所以 $\omega\in A$ 且 $\omega\in C$ 或 $\omega\in B$ 且 $\omega\in C$，即 $\omega\in AC$ 或 $\omega\in BC$，亦即 $\omega\in(AC)\cup(BC)$。因此

$$(A\cup B)C\subset(AC)\cup(BC) \tag{2-12}$$

由式（2-11）和式（2-12）可得结论 2）。

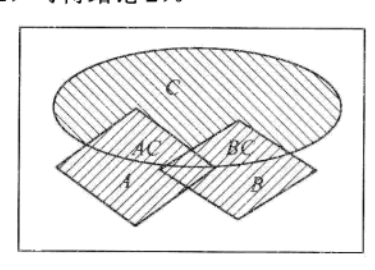


图 2-6　$(A\cup B)C=(AC)\cup(BC)$ 的示意图

利用事件运算的对偶律和结论 2）得

$$\begin{aligned}(AB)\cup C&=\overline{\left(\overline{AB}\right)\overline{C}}=\overline{\left(\overline{A}\cup\overline{B}\right)\overline{C}}\\&=\overline{\left(\overline{A}\overline{C}\right)\cup\left(\overline{B}\overline{C}\right)}=\overline{\left(\overline{A}\overline{C}\right)}\ \overline{\left(\overline{B}\overline{C}\right)}\\&=(A\cup C)(B\cup C)\end{aligned}$$

即结论 3）成立。

事件 $(A_1\cup A_2)\cup A_3$ 和 $A_1\cup(A_2\cup A_3)$ 相等，它们都是表示把 A_1、A_2 和 A_3 中的所有样本点合在一起所构成的事件，即表示“事件 A_1、A_2 和 A_3 中至少有一个事件发生”。为简单起见，用 $A_1\cup A_2\cup A_3$ 表示该事件。

一般地，事件的并和交的运算可以推广到任意多个事件的情况。对于事件列$\{A_n\}$，用$\bigcup_{k=1}^{n} A_k$表示"在事件$A_1, A_2, \cdots, A_n$中至少有一个事件发生"，并称它为该事件的并；用$\bigcup_{k=1}^{\infty} A_k$表示"在事件列$\{A_n\}$中至少有一个事件发生"，并称它为该事件列的并；用$\bigcap_{k=1}^{n} A_k$表示"事件$A_1, A_2, \cdots, A_n$同时发生"，并称它为该事件的交；用$\bigcap_{k=1}^{\infty} A_k$表示"在事件列$\{A_n\}$中所有事件都同时发生"，并称它为该事件列的交。

【例 2-15】 在例 2-12 中，用A_n表示"该网站在 1h 之内被点击 n 次"，则$\Omega = \bigcup_{k=0}^{\infty} A_n$。用$B_n$表示"该网站在 1h 之内被点击次数大于或等于 n"，则$B_n = \bigcup_{k=n}^{\infty} A_k$。进一步，$\bigcap_{n=1}^{\infty} B_n$表示"该网站在 1h 内被点击无穷多次"。

借助于事件的运算，可以把复杂的概率计算问题转化为简单的概率计算问题，在随后的学习和研究过程中将会体会到这一点。

2.3 条件概率与事件的独立性

2.3.1 条件概率

前面讨论随机事件 B 的概率都是相对于某组确定的条件 S 而言的。$P(S)$就是在条件组 S 的实现之下，事件 B 发生的概率。有时除了这组基本条件 S 之外，还要提出附加的条件，也就是要求"在事件 A 已经发生的前提下，事件 B 发生的概率"。这就是所谓的条件概率问题，记为$P(B|A)$。

下面首先研究一个简单的示例。

【例 2-16】 一盒产品共 10 只，其中有 3 只次品。现无放回地从中抽取两次，每次任取一只。

1）第二次取到次品的概率是多少？

2）第一次取到次品后，第二次取到次品的概率是多少？

解：1）令$A=$"第一次取到次品"，$B=$"第二次取到次品"，显然$P(A)=\frac{3}{10}$，那么$P(B)$等于多少？由于"第二次取到次品"这一事件 B 对第一次取到什么产品没有限制或假设，因此回答

$$P(B)=\frac{3}{9} \text{或} P(B)=\frac{2}{9}$$

都是没有根据的，但是这个问题可以用古典概型与概率的加法公式求解。

由于

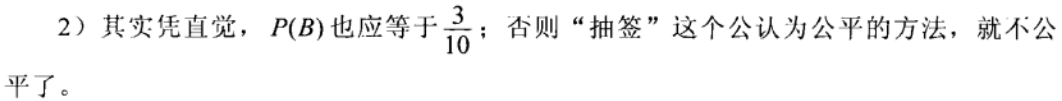
$$B = AB + \overline{A}B，\ (AB)(\overline{A}B) = \varnothing$$

所以

$$P(B)=P(AB)+P(\overline{A}B)=\frac{C_3^1C_2^1}{A_{10}^2}+\frac{C_7^1C_3^1}{A_{10}^2}$$

$$=\frac{3\times 2}{10\times 9}+\frac{7\times 3}{10\times 9}=\frac{3}{10}$$





2）其实凭直觉，$P(B)$ 也应等于 $\frac{3}{10}$；否则“抽签”这个公认为公平的方法，就不公平了。

至于 $P(B|A)$ 是在一定条件下，又附加了一个 A 已经发生的条件，事件 B 发生的概率。在上例中 $P(B|A)\neq P(B)$，即 $P(B|A)$ 与 $P(B)$ 是有区别的，所以称其为条件概率。一般地，有如下的定义。

定义 2-6 如果 A 及 B 是条件组 S 下的两个随机事件，且 $P(A)\neq 0$，则称在事件 A 发生的前提下事件 B 发生的概率为条件概率，记为 $P(B|A)$。

2.3.2 乘法公式

由上述内容可知，条件概率 $P(B|A)$ 与事件 B 的原概率 $P(B)$ 在一般的情况下是不相等的，那么它们之间有什么关系呢？人类从长期的大量社会实践中总结出了它们之间具有如下的普遍规律，即

$$P(B|A)=\frac{P(AB)}{P(A)}\ (P(A)\neq 0) \tag{2-13}$$

式（2-13）揭示了概率 $P(A)$、$P(AB)$ 与条件概率 $P(B|A)$ 之间的关系。通常，可以从如下的两个方面来利用这一关系。

1）已知 $P(A)$、$P(AB)$ 来求得 $P(B|A)$。

2）已知 $P(A)$，$P(B|A)$ 来求得 $P(AB)$。

在后一种情况下，为了方便起见，还可将式（2-13）改写为

$$P(AB)=P(A)P(B|A) \tag{2-14}$$

式（2-14）称为概率的乘法公式。

【例 2-17】 盒中有 5 个乒乓球，其中 3 个新球，2 个旧球。现无放回地取两次，每次任取一球，求第一次取到新球后，第二次取到新球的概率。

解： 令 $A=$“第一次取到新球”，$B=$“第二次取到新球”，则有 $AB=$“第一、二次都取到新球”，且

$$P(A)=\frac{3}{5}，\ P(AB)=\frac{C_3^1\cdot C_2^1}{A_5^2}=\frac{3\times 2}{5\times 4}=\frac{3}{10}$$

利用式（2-13）即得所求事件的概率为

$$P(B|A)=\frac{P(AB)}{P(A)}=\frac{3}{10}\Big/\frac{3}{5}=\frac{1}{2}$$

【例 2-18】 甲乙两厂共生产了 1000 个零件，其中 300 个是乙厂生产的。而在这 300 个零件中有 189 个是标准品。现从 1000 个零件中任取一个，问是乙厂生产的标准品的概率是

多少？

解： 令 $A=$“取出的是乙厂生产的标准品”。

则有

$AB=$“取出的是甲乙两厂生产的标准品”。

又由于

$$P(A)=\frac{300}{1000}=0.3\text{，}\quad P(B\mid A)=\frac{189}{300}=0.63$$

所以利用式（2-14）得

$$P(AB)=P(A)P(B\mid A)=0.3\times 0.63=0.189$$

2.3.3 独立性

1．两个事件的独立性

在给出独立的概念之前，先看一个简单的例子。

【例 2-19】 盒子中共有 5 个乒乓球，其中 3 个新球，2 个旧球。现有放回地抽取两次，每次任取 1 球。如果记 $A=$“第一次取到新球”，$B=$“第二次取到新球”，则有

$$P(B\mid A)=P(B\mid \overline{A})=P(B)\text{。}$$

解： 显然有

$\overline{A}=$“第一次取到旧球”，

且

$$P(A)=P(B)=\frac{3}{5}\text{，}\quad P(\overline{A})=\frac{2}{5}$$

又由于

$$P(AB)=\frac{\mathrm{C}_3^1\cdot\mathrm{C}_3^1}{\mathrm{U}_5^2}=\frac{3^2}{5^2}=\frac{9}{25}\text{，}\quad P(\overline{A}B)=\frac{\mathrm{C}_3^1\cdot\mathrm{C}_2^1}{\mathrm{U}_5^2}=\frac{3\times 2}{5^2}=\frac{6}{25}$$

所以

$$P(B\mid A)=\frac{P(AB)}{P(\overline{A})}=\frac{9}{25}\Big/\frac{3}{5}=\frac{3}{5}$$

$$P(B\mid \overline{A})=\frac{P(\overline{A}B)}{P(A)}=\frac{6}{25}\Big/\frac{2}{5}=\frac{3}{5}$$

因此

$$P(B\mid A)=P(B\mid \overline{A})=P(B)$$

例 2-19 说明，在某些情况下，事件 A 的发生或不发生均不影响事件 B 发生的概率，且可以说明

$$P(B\mid A)=P(B)\Leftrightarrow P(AB)=P(A)P(B)$$

其中 $P(A)\neq 0$（证明这里从略）。由此便可引出随机事件的独立性概念。

定义 2-7 如果随机事件 A 及 B 满足

$$P(AB)=P(A)P(B) \tag{2-15}$$

则称 A 与 B 为相互独立的随机事件。



定理 2-3 如果 4 对事件：

1）A 与 B。

2）A 与 $\overline{B}$。

3）$\overline{A}$ 与 B。

4）$\overline{A}$ 与 $\overline{B}$。

即上述 4 对事件或者都相互独立，或者都不独立。

证明：只需证明如果 A 与 B 独立，则 $\overline{A}$ 与 B 也独立，其余读者自行证明。

因为

$$B = AB + \overline{A}B，\quad (AB)(\overline{A}B) = \varnothing$$

所以

$$P(B) = P(AB) + P(\overline{A}B)$$

又因为 A 与 B 相互独立，即

$$P(AB) = P(A)P(B)$$

所以

$$\begin{aligned} P(\overline{A}B) &= P(B) - P(AB) = P(B) - P(A)P(B) \\ &= [1 - P(A)]P(B) = P(\overline{A})P(B) \end{aligned}$$

即 $\overline{A}$ 与 B 相互独立。

【例 2-20】 两射手彼此独立地向同一目标射击。设甲射中的概率为 0.9，乙射中的概率为 0.8，求目标被击中的概率。

解：令 A = “甲射中目标”；B = “乙射中目标”，C = “目标被击中”，则有

$$C = A + B$$

方法一：

$$\begin{aligned} P(C) &= P(A+B) = P(A) + P(B) - P(AB) \\ &= P(A) + P(B) - P(A)P(B) \\ &= 0.9 + 0.8 - 0.9 \times 0.8 = 0.98 \end{aligned}$$

方法二：

因为

$$P(\overline{C}) = P(\overline{A+B}) = P(\overline{A}\overline{B}) = P(\overline{A})P(\overline{B})$$

所以

$$\begin{aligned} P(C) &= 1 - P(\overline{C}) = 1 - P(\overline{A})P(\overline{B}) \\ &= 1 - [1 - P(A)][1 - P(B)] \\ &= 1 - (1 - 0.9)(1 - 0.8) = 1 - 0.02 = 0.98 \end{aligned}$$

2. n 个事件的独立性

定义 2-8 如果事件 A，B，C 满足

$$P(AB) = P(A)P(B)，\quad P(BC) = P(B)P(C)$$

$$P(CA) = P(C)P(A)，\quad P(ABC) = P(A)P(B)P(C)$$

则称 A，B，C 是相互独立的事件。

定义 2-8 给出了 3 个事件相互独立的概念。一般地，对于 n 个事件 $A_1, A_2, \cdots, A_n$ 的相互独立性，有如下的定义。

定义 2-9 如果事件组 $A_1, A_2, \cdots, A_n$ 对于任取的正整数 k（$2 \leqslant k \leqslant n$）和任意的 $1 \leqslant i_1 \leqslant i_2 \leqslant i_k \leqslant n$，均满足等式：

$$P(A_{i_1}, A_{i_2}, \cdots, A_{i_k}) = P(A_{i_1})P(A_{i_2})\cdots P(A_{i_k}) \tag{2-16}$$

则称事件组 $A_1, A_2, \cdots, A_n$ 是相互独立的。

如何判断一些事件是否相互独立呢？在很多情况下，并不需要利用式（2-16）进行复杂的计算，而是根据对事件本质的分析即可知道。

显然，当事件组 $A_1, A_2, \cdots, A_n$ 相互独立时，有

$$P(A_1, A_2, \cdots, A_n) = P(A_1)P(A_2)\cdots P(A_n) \tag{2-17}$$

【例 2-21】 设每支步枪射中飞机的概率为 0.004。

1）求 250 支步枪同时射击时击中飞机的概率。

2）要使击中飞机的概率达到 99%，至少需要多少支步枪？

解： 令 $A =$ “一支步枪射中飞机”。

$B =$ “250 支步枪同时射击时击中飞机”。

$A_i =$ “第 i 支步枪射中飞机”。

则有

$$B = A_1 + A_2 + \cdots + A_{250}, \quad \overline{B} = \overline{A}_1\overline{A}_2\cdots\overline{A}_{250}$$

$$P(A) = P(A_i) = 0.004, \quad P(\overline{A}) = P(\overline{A}_i) = 0.996$$

1）由于

$$P(\overline{B}) = P\left(\overline{A}_1\overline{A}_2\cdots\overline{A}_{250}\right) = P(\overline{A}_1)P(\overline{A}_2)\cdots P(\overline{A}_{250})$$
$$= [P(\overline{A})]^{250} \approx 0.37$$

因此所求事件 B 的概率为

$$P(B) = 1 - P(\overline{B}) = 1 - [P(\overline{A})]^{250}$$
$$= 1 - (0.996)^{250} \approx 0.63$$

2）由 1）可知，n 支步枪同时射击飞机时击中飞机的概率应为

$$1 - [P(\overline{A})]^n = 1 - (0.996)^n$$

从而由题意知

$$1 - (0.996)^n \geqslant 0.99 \text{ 或 } (0.996)^n \leqslant 0.01$$

即有

$$n\ln(0.996) \leqslant \ln(0.01) \text{ 或 } n \geqslant 1150$$

因此所需步枪数至少为 1150 支。

【例 2-22】 一射手每次击中某目标的概率为 $p(0 < p < 1)$，现独立地向该目标射击了 n 次，求其击中目标的概率。

解： 令 $A =$ “击中目标”，$A_i =$ “第 i 次射中目标”，

则有

$$A = A_1 + A_2 + \cdots + A_n,\quad \overline{A} = \overline{A}_1 + \overline{A}_2 + \cdots + \overline{A}_n$$

又由于

$$P(A_i) = p,\quad P(\overline{A}_i) = 1 - p(i = 1,2,\cdots,n)$$



因此，所求事件的概率为

$$\begin{aligned} P(A) &= 1 - P(\overline{A}) = 1 - P(\overline{A}_1\overline{A}_2\cdots\overline{A}_n) \\ &= 1 - P(\overline{A}_1)P(\overline{A}_2)\cdots P(\overline{A}_n) = 1 - (1-p)^n \end{aligned}$$

2.4 概率空间

20 世纪前，还没有提出概率论的公理化体系，主要的研究范围为古典概型和几何概型，这限制了概率论的发展。数学家科尔莫戈罗夫于 1934 年出版的《概率论的基本概念》标志着概率论的公理化体系的建立。

2.4.1 基本概念

概率论的主要任务是研究概率所共有的性质，这里的概率是指研究者所感兴趣的事件的概率。例如，在掷骰子实验中，如果只关心掷出的点数是否为偶数，那么所关心的事件就是 $A = \{2,4,6\}$ 是否出现，而不关心事件$\{1\}$、$\{3\}$等。

用 $\wp$ 表示所关心的事件全体。一个自然的想法是仅把 $\wp$ 中的事件作为研究范围，以便集中精力研究所关心的事件出现的概率。在很多时候，利用事件的运算可以简化所关心事件的概率计算，因此应该要求研究范围对事件（有限个）的运算封闭，即研究范围内的任何两个（有限个）事件的运算结果还应该在研究范围之中。

在掷骰子实验中，如果关心掷出的点数是否小于或等于 5，则所关心的事件全体为 $\wp = \{A\}$，其中 $A = \{1,2,3,4,5\}$，显然 $\wp$ 对事件的运算不封闭，如 $\overline{A} = 6 \notin \wp$，而在某些情况下事件 $\overline{A}$ 的概率更容易得到。要使得研究的范围对事件的运算封闭，这个范围应该是

$$\wp = \{\varnothing, A, \overline{A}, \Omega\}$$

或者是样本空间 Ω 的一切子集全体

$$\overline{\wp} = \{B \mid B \subset \Omega\}$$

等。这里 $\Omega = \{1,2,3,4,5,6\}$ 是必然事件，显然 $\wp$ 中仅有 4 个事件，而 $\overline{\wp}$ 中则有 $2^6 = 64$ 个事件。也就是说，在这种情况下，$\overline{\wp}$ 中包含了更多的我们所不感兴趣的事件。

若 $\wp$ 是由一些事件构成的集合，则它对事件的有限次运算是封装的，且 $\Omega = \wp$，则称 $\wp$ 为 Ω 上的 σ 代数，即任何 σ 代数都被它包含。

【例 2-23】 考虑一个随机现象可能会出现的事件 A。在相同的条件下重复观测该现象 n 次，用 $n(A)$ 表示 n 次观测中 A 出现的次数，称

$$F(A) = \frac{n(A)}{n}$$

为事件 A 发生的频率。

定理 2-4 频率 F 具有如下基本性质：

1）非负性：$F(A) \geqslant 0$，$\forall A$。

2）规范性：$F(\Omega)=1$，$F(\varnothing)=0$。

3）可加性：若事件 A 与 B 不相容，则 $F(A \cup B)=F(A)+F(B)$。

证明：非负性和规范性显然成立。注意到当 $A \cap B=\varnothing$ 时，有 $n(A \cup B)=n(A)+n(B)$，即可加性。

在一定的条件下，当实验的次数 $n \to \infty$ 时，$F(A)$ 的“极限”存在，这时称它为事件 A 的概率，记为 $P(A)$。

由于极限具有可加性和保序性，进而由频率的 3 条性质可以推出概率应该具有的相应性质：

1）非负性：$F(A) \geqslant 0$。

2）规范性：$F(\Omega)=1$，$F(\varnothing)=0$。

3）可加性：若事件 A 与 B 不相容，则 $F(A \cup B)=F(A)+F(B)$。

2.4.2 概率空间

假设 σ 代数 $\wp$ 是研究范围，对于 $\wp$ 中的每一个事件，都有一个概率值与之对应，即概率从 $\wp$ 到实数集合的一种映射。通常，称从 $\wp$ 到实数集合的映射为集函数。按照这种观点，概率是一个集函数，反之却不真。例如，把所有的事件都映射成 -1 的函数显然不是一个概率。一个自然的问题是，作为概率的集函数应该有什么最基本的性质？什么样的性质可以作为概率理论研究的理论基础？

定义 2-10（概率的基本公理） 设 $\wp$ 为 Ω 上的 σ 代数，如果定义在 $\wp$ 上的集函数 $\Re(\cdot)$ 满足如下条件：

1）非负性：$\forall A \in \wp$，有 $\Re(A) \geqslant 0$。

2）规范性：$\Re(\Omega)=1$。

3）可列可加性：对于两两不相容的事件列 $\{A_n\} \subset \wp$，有

$$\Re\left(\bigcup_{n=1}^{\infty} A_n\right)=\sum_{n=1}^{\infty} \Re(A_n)$$

则称 $\Re$ 为 $\wp$ 上的概率测度，简称为概率；称 $\Re(A)$ 为事件 A 的概率；称 $(\Omega, \wp, \Re)$ 为概率空间。

注意：比较定理 2-4 中列出的频率的 3 条性质，不难看出，可把非负性、规范性和可列可加性作为概率公理的背景。

【例 2-24】 对于样本空间 Ω 的一个子集 A，取 $\wp=\{\varnothing, A, \bar{A}, \Omega\}$。定义

$$P(A) \triangleq p，\quad P(\bar{A}) \triangleq q，\quad P(\phi) \triangleq 0，\quad P(\Omega) \triangleq 1$$

其中，$0<p<1$，$q=1-p$，则称 $(\Omega, \wp, \Re)$ 为伯努利概率空间。

注意：在伯努利概率空间中，可以认为样本空间只有两个样本点 A 和 $\bar{A}$，即 $\Omega=\{A, \bar{A}\}$。若随机实验只有“成功”与“失败”两个可能的实验结果，则可以用伯努利概率空间来描述这个随机实验。只需把 A 理解为成功，把 $\bar{A}$ 理解为失败即可。此时，称 p 为成功概率，称 q 为失败概率。

伯努利概率空间应用很广，比如可以应用于掷硬币、考察产品是否合格、射击是否击中目标等问题。一般地，凡是只有两个互斥结果的随机现象都可以用伯努利概率空间来描述。

【例 2-25】 样本空间 $\Omega=\{\omega_1,\omega_2,\cdots,\omega_n\}$，$\wp$ 由 Ω 的一切子集所构成。如果样本点 ω_i 出现的概率为 p_i，$1\leqslant i\leqslant n$，定义

$$P(A)\triangleq\sum_{i:\ \omega_i\in A}p_i\text{，}\ \forall A\in\wp \tag{2-18}$$

则称 $(\Omega,\wp,\Re)$ 为有限概率空间。特别地，当 $p_i=1/n$ 时，称这个概率空间为古典概率空间，相应的概率称为古典概率。

注意：1）显然 $\sum_{i=1}^{n}p_i=1$。

2）式（2-18）等号的右端表示 A 中的样本点所对应的概率之和。例如，在掷骰子实验中，用 A 表示掷出"偶数点"，p_i 表示掷出"i 点"的概率，即

$$A=\{2,4,6\}\text{，}\ P(\{i\})=p_i\text{，}\ 1\leqslant i\leqslant 6$$

则

$$\sum_{i:\ \omega_i\in A}p_i=p_2+p_4+p_6$$

【例 2-26】 样本空间 $\Omega=\{\omega_1,\omega_2,\cdots,\omega_n\}$，$\wp$ 由 Ω 的一切子集所构成。如果样本点 ω_i 出现的概率为 p_i，$1\leqslant i\leqslant n$ 定义

$$P(A)\triangleq\sum_{i:\ \omega_i\in A}p_i\text{，}\ \forall A\in\wp \tag{2-19}$$

则称 $(\Omega,\wp,\Re)$ 为离散概率空间。

注意：这里，事件 A 的概率表达式和有限概率空间中的相同。不同的是，A 中可能包含无限多个样本点，因此定义的右端可能是一个无穷和。

在例 2-12 中，样本空间 Ω 由所有非负整数构成。令

$$\wp=\{A\,|\,A\in\Omega\}\text{，}\ p_i=\frac{1}{i!}\mathrm{e}^{-1} \tag{2-20}$$

其中，i 为非负整数，并按式（2-20）定义 $\Re$，则 $(\Omega,\wp,\Re)$ 为离散概率空间。特别地，有

$$P\text{（“该网站被点击奇数次”）}=\sum_{i=0}^{\infty}\frac{1}{(2i+1)!}\mathrm{e}^{-1}$$

等号右边为无穷和。

【例 2-27】 对于 n 维欧氏空间 P^n 中的一个区域 A，用 $m(A)$ 表示其体积。设样本空间 Ω 为 P^n 中的一个区域，满足 $0<m(\Omega)<\infty$。用 $\wp$ 表示 Ω 的可求体积的子区域的全体。定义

$$P(A)\triangleq\frac{m(A)}{m(\Omega)} \tag{2-21}$$

则称 $(\Omega,\wp,\Re)$ 为几何概率空间，相应的概率称为几何概率。

注意：在一维欧氏空间（即数轴）中的“体积”就是长度；在二维欧氏空间（即二维平面）中的“体积”就是面积。在几何概率空间中，求概率转化为求体积（长度或面积）。

【例 2-28】 甲乙两人约定在 12～13 点之间（不包括 13 点）的任何一个时刻去公园会面，规定先到者仅等候 30min。试求事件 A={甲乙能见面}的概率。

解：假定甲和乙都在 12 点以后到达会面地点。用 x 和 y 分别表示甲和乙到达会面地点时距离 12 点的分钟数，则样本空间

$$\Omega = \{(x,y) \mid 0 \leqslant x, y < 60\}$$

为二维欧氏空间中的多边形，如图 2-7a 所示，其面积 $m(\Omega) = 3600$。两个人能见面的条件是 $|x-y| \leqslant 30$，所以

$$A = \{(x,y) \in \Omega \mid |x-y| \leqslant 30\}$$

为二维欧氏空间中的多边形，即图 2-7b 中的灰色区域。由于两个人到达时刻的任意性，因此可以用几何概率来计算事件 A 的概率。现在 $m(A) = 60^2 - 30^2 = 2700$，所以两个人能见面的概率为 3/4。

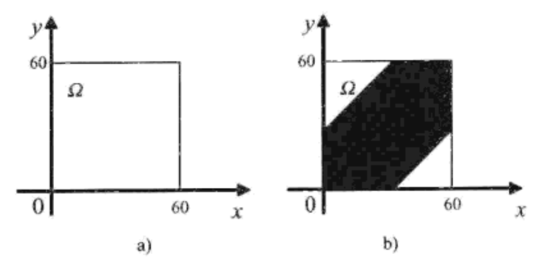


图 2-7 样本空间与事件 A

a) 正方形区域（Ω） b) 灰色区域（A）

注意：这里“两个人到达时刻的任意性”的含义如下：对于矩形 $[0,60) \times [0,60)$ 中的任何区域 A，$A(x,y) \in A$ 的概率仅和该区域的面积有关。也就是说，如果 $B \subset [0,60) \times [0,60)$ 的面积和 A 的面积相等，则

$$P((x,y) \in A) = P((x,y) \in B)$$

从概率定义的 3 条公理出发，可以推出许多概率的性质，这些性质可以用来帮助计算复杂事件的概率。下面介绍几个常用的简单的概率性质。

定理 2-5 设 $(\Omega, \wp, \Re)$ 为概率空间，则概率有如下性质：

1）$P(\varnothing) = 0$。

2）有限可加性：任意两两不相容的事件 $A_1, A_2, \cdots, A_n \in \wp$，则

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i)$$

3）对于任何事件 $A \in \wp$，$0 \leqslant P(A) \leqslant 1$。

4）可减性：对任何事件 $A \subset B$，有

$$P(B-A) \leqslant P(B) - P(A)$$

5）单调性：对任何事件 $A\subset B$，有

$$P(A)\leqslant P(B)$$

6）对于任何事件 A，有 $P(\overline{A})=1-P(A)$。

7）加法公式：对于任何事件 A 和 B，有

$$P(A\cup B)\leqslant P(A)+P(B)-P(AB)$$



证明：对于任何正整数 n，取 $A_n=\varnothing$，则 $\{A_n\}$ 为两两不相容的事件列，且 $\varnothing=\bigcup\limits_{k=1}^{\infty}A_k$。这样，由概率的可列可加性，有

$$P(\varnothing)=P\left(\bigcup_{k=1}^{\infty}A_k\right)=\sum_{k=1}^{\infty}P(A_k)=\sum_{k=1}^{\infty}P(\varnothing)$$

因此 $P(\varnothing)=0$，即性质 1）成立。

若 $A_1,A_2,\cdots,A_n$ 两两不相容，可取 $A_{n+k}=\varnothing$，$k\geqslant1$，则 $\{A_k\}$ 为两两不相容的事件列，且 $\bigcup\limits_{k=1}^{n}A_k=\bigcup\limits_{k=1}^{\infty}A_k$，由概率的可列可加性和性质 1）得

$$P\left(\bigcup_{k=1}^{n}A_k\right)=P\left(\bigcup_{k=1}^{\infty}A_k\right)=\sum_{k=1}^{\infty}P(A_k)$$

$$=\sum_{k=1}^{n}P(A_k)+\sum_{k=n+1}^{\infty}P(A_k)=\sum_{k=1}^{n}P(A_k)$$

即性质 2）成立。

对于任何事件 $A\in F$，有 $\Omega=A\cup\overline{A}$，且 A 与 $\overline{A}$ 互不相容，由概率的非负性、规范性和有限可加性得

$$1=P(\Omega)=P(A)+P(\overline{A})\geqslant P(A)\geqslant 0$$

即性质 3）成立。

当 $A\subset B$ 时，有 $B=A\cup(B-A)$，再注意到 $A(B-A)=\varnothing$，由性质 2）得

$$P(B)=P(A)+P(B-A)$$

再由性质 3）得性质 4）。

由概率的可减性和非负性得 $P(B)=P(A)+P(B-A)\geqslant P(A)$，即性质 5）成立。

注意到 $A\subset\Omega$，$\overline{A}=\Omega-A$，利用概率的规范性和有限可加性得

$$1=P(\Omega)=P(A)+P(\Omega-A)=P(A)+P(\overline{A})$$

即性质 6）成立。

由有限可加性易得性质 7）。

定理 2-6 若 $(\Omega,\wp,\Re)$ 为古典概率空间，对于任何事件 $A\in\wp$，$n(A)$ 表示事件 A 中样本点的个数，则

$$P(A)=\frac{n(A)}{n(\Omega)},\quad A\in\wp \tag{2-22}$$

注意：在古典概率空间中，概率的计算转化为计数，称式（2-22）为古典概率计算

公式。

证明：设 $\Omega=\{\omega_1,\omega_2,\cdots,\omega_m\}$，则

$$P(\{\omega_k\})=\frac{1}{m},\quad k=1,2,\cdots,m \tag{2-23}$$

对于任何事件 $A\in\wp$，记 $k=n(A)$，则存在 $\omega_{n_i}\in\Omega$，$1\leqslant i\leqslant k$，使得

$$A=\{\omega_{n_1},\omega_{n_2},\cdots,\omega_{n_k}\}=\{\omega_{n_1}\}\cup\cdots\cup\{\omega_{n_k}\}$$

注意到 k 个事件 $\{\omega_{n_1}\}\cdots\{\omega_{n_k}\}$ 两两不相容，由概率的有限可加性和式（2-23）知

$$P(A)=P(\{\omega_1\})+\cdots+P(\{\omega_k\})=\frac{k}{m}$$

再注意到 $n(\Omega)=m$ 和 $k=n(A)$，即可得结论。

【例 2-29】 试证明投掷一枚质地均匀的骰子的结果可以用古典概率空间来描述，并给出出现任何一个结果 A 的概率计算公式。

证明：在投掷均匀骰子的实验中，样本空间 $\Omega=\{1,2,3,4,5,6\}$，事件类

$$\wp=\{A\mid A\subset\Omega\}$$

由于骰子的质地均匀，所以出现事件 $\{i\}$ 的概率为 $p_i=1/6$。因此概率空间 $(\Omega,\wp,\Re)$ 为古典概率空间，其中概率由式（2-19）定义。根据古典概率计算公式（2-22），任何一个结果 $A\in\wp$ 的概率为

$$P(A)=\frac{A\text{中样本点的个数}}{6}$$

2.5 总体样本

在前面内容中讲述了概率论的初步知识，随后各章将讲述数理统计。数理统计是具有广泛应用的一个数学分支，它以概率论为理论基础，根据实验或观察得到的数据来研究随机现象，对研究对象的某些规律性作出合理的估计和判断。数理统计学的重要分支有统计推断、多元统计分析和试验设计等。其具体方法很多，应用相当广泛，已成为科学研究及生产、经济等部门进行有效研究工作中必不可少的数学工具。

总体、个体、样本（子样）是数理统计学中 3 个基本的术语。在数理统计中，常关注研究对象的某项数量指标，将研究对象的某项数量指标的值的全体称为总体，总体中的每个元素称为个体，每个个体是一个实数。例如，某学校女生的身高的全体是一个总体，每一个女生的身高是一个个体；某地在某季度内每天的日平均气温的全体是一个总体，某天的日平均气温是一个个体。

2.5.1 总体与样本的基础

1．示例

【例 2-30】 某市在职职工有 100 万，那么怎样得到职工的年收入情况呢？

【例 2-31】 由于种种因素的影响，灯泡厂生产出来的灯泡的寿命是不同的。为了判断所生产灯泡的质量，怎样去估计某天所生产的灯泡的平均寿命，以及使用时数长短的相差



程度？

【例 2-32】 人的血型在医疗上（输血）无疑是重要的。假如要掌握华北地区人们的血型分布（按 A、B、AB、O 型分组），那么如何去获得这种资料？

事实上，某市职工年收入情况、灯泡厂灯泡的平均寿命及各小时寿命的比例、华北地区人们血型分布都是客观存在的，只是在研究之前不知道具体细节。一般地，研究的总体，即研究对象的某项数量指标 X，它的取值在客观上有一定的分布，X 是一个随机变量。对总体的研究，就是对相应的随机变量 X 的分布的研究。因此，X 的分布函数和数字特征分别成为总体的分布函数和数字特征。今后将不区分总体和相应的随机变量。

那么如何获得总体的分布情况呢？

要将一个总体的情况了解得十分清晰，初看起来，最理想的办法是对每个个体逐个进行观察或实验，但实际上这样做往往是不现实的。例如，要研究灯泡寿命，要对每个灯泡逐个观察寿命，由于寿命实验是破坏性的，一旦获得试验的所有结果，这批灯泡也就全部烧毁了。因此灯泡的寿命不适宜采用普查的方法。对某城市 100 万职工年收入情况的了解和华北地区人们的血型分布也不宜采用普查的方法，因为除客观因素的制约（如血型普查必须得到每人抽取血样的认可）外，投入人力物力太多，耗时太多，也是一个制约因素。总之，对总体情况的了解若遇到观察具有破坏性（如灯泡寿命）和基数太大，投入人力物力和时间等太多，则一般不宜采用普遍逐个观察的方法。那么采用什么方法来解决问题呢？一个很重要的方法是随机抽样法。从整批灯泡中抽取一些灯泡做寿命试验，并记录其结果，然后根据这些数据来推断整批灯泡的寿命情况。对于职工年收入的情况，从职工中抽取一部分，记录他们的年收入数据，然后借此数据推断整体职工年收入的分布情况。对血型分布情况，采用在不同人群中抽取一部分人，记载其血型，并根据记载情况来推断整个地区的血型分布。

对于类似上述例子的问题，都是从总体中抽取部分个体进行观察，然后根据所得数据来推断总体的性质。

2. 概念

在一个总体 X 中，抽取 n 个个体 $X_1,X_2,\cdots,X_n$，这 n 个个体 $X_1,X_2,\cdots,X_n$ 称为总体 X 的一个容量为 n 的样本（或称子样）。

所谓从总体抽取一个个体，就是对总体 X 进行一次观察（即进行一次试验），并记录其结果。在相同的条件下对总体 X 进行 n 次重复的、独立的观察，将 n 次观察结果按照试验的次序记为 $X_1,X_2,\cdots,X_n$，由于 $X_1,X_2,\cdots,X_n$ 是对随机变量 X 观察的结果，且每次观察是在相同的条件下独立进行的，所以可认为 $X_1,X_2,\cdots,X_n$ 是相互独立的，且都是与 X 具有相同分布的随机变量。这样得到的 $X_1,X_2,\cdots,X_n$ 称为来自总体 X 的一个简单随机样本，n 称为样本容量。

当 n 次观察已经完成，就得到一组实数 $x_1,x_2,\cdots,x_n$，它们依次是随机变量 $X_1,X_2,\cdots,X_n$ 的观察值，称为样本值（或样本观察值）。

定义 2-10 设 X 是具有分布函数 F 的随机变量，若 $X_1,X_2,\cdots,X_n$ 是具有同一分布函数 F 且相互独立的随机变量，则称 $X_1,X_2,\cdots,X_n$ 为从分布函数 F（或总体 F、总体 X）得到的容量为 n 的简单随机样本，简称样本。它们的观察值 $x_1,x_2,\cdots,x_n$ 称为样本值，又称为 X 的 n

个独立的观察值。

以后如无特别说明，所提到的样本都是指简单随机样本。

2.5.2 分布定理

定理 2-7 若 $X_1,X_2,\cdots,X_n$ 是来自总体 X 的一个样本，$F(x)$ 是 X 的分布函数，$p(x)$ 是 X 的密度函数，则 $X_1,X_2,\cdots,X_n$ 的联合分布函数 $F^*(x_1,x_2,\cdots,x_n)$ 和联合概率密度 $p^*(x_1,x_2,\cdots,x_n)$ 分别为

$$F^*(x_1,x_2,\cdots,x_n)=\prod_{i=1}^{n}F(x_i)$$
$$p^*(x_1,x_2,\cdots,x_n)=\prod_{i=1}^{n}p(x_i) \tag{2-24}$$

还需要指出，在实际抽样中，对于有限个个体组成的总体，采用放回抽样就能得到简单随机样本。有时放回抽样使用起来不方便，当个体总数 N 比要得到的样本容量 n 大得多时，在实际中，可将不放回抽样近似地视为放回抽样（即简单随机样本）处理。

2.6 统计量与抽样分布

2.6.1 统计量

统计量是统计分析的基本工具。

统计量是指样本的不含其他未知参数的函数。统计量概念的要点是“不含其他未知参数”，即只要给定样本数据，则统计量的函数值（统计量的观测值）就能够唯一地确定下来。

统计分析技术在一定程度上可以说是统计量的构造技术。学习过程中要高度重视针对某种特定的问题是如何构造相关统计量的。本小节仅讲解几类基本的统计量，这是在特定的问题中构造相关统计量的基础材料。

（1）样本矩

样本矩是最基本、常用的一类统计量，主要包括如下两种：

1）样本 k 阶（原点）矩。设 $X_1,X_2,\cdots,X_n \sim X$，则称

$$A_k=\frac{1}{n}\sum_{i=1}^{n}X_i^k\ ,\quad k=1,2,\cdots \tag{2-25}$$

为变量 X 的样本 k 阶（原点）矩。A_k 的观测值记为 μ_k。

特别地，样本的 1 阶矩

$$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i \tag{2-26}$$

称为样本均值。它是最重要的统计量之一，反映了变量 X 取值集中程度的信息。$\overline{X}$ 的观测值用 $\overline{x}$ 表示。

2）样本 k 阶中心矩。设 $X_1,X_2,\cdots,X_n \sim X$，则称

$$B_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^k \text{，} \quad k = 1,2,\cdots \tag{2-27}$$

为变量 X 的样本 k 阶中心矩。B_k 的观测值记为 v_k。



特别地，样本的 2 阶中心矩

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \tag{2-28}$$

称为样本方差。它也是最重要的统计量之一，反映了变量 X 取值分散程度的信息。S^2 的观测值用 s^2 表示。

值得注意的是，在实际应用中，常用样本的修正方差

$$S^{*2} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \tag{2-29}$$

替代 S^2（以下若无特别说明，“样本方差”一词均指样本的修正方差，仍记为 S^2）。样本方差的算术根称为样本标准差，记为 S，即

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

（2）顺序统计量

顺序统计量是另一类最基本、常用的统计量。

设 $X_1, X_2, \cdots, X_n$ i.i.d $\sim X$，$(x_1, x_2, \cdots, x_n)$ 是 $(X_1, X_2, \cdots, X_n)$ 的任意一次观测值。记为 $x_{(1)}, x_{(2)}, \cdots, x_{(n)}$ 是 $x_1, x_2, \cdots, x_n$ 的一个排列，并且 $x_{(1)} \leqslant x_{(2)} \leqslant \cdots \leqslant x_{(n)}$。若令 n 维随机向量 $(X_{(1)}, X_{(2)}, \cdots, X_{(n)})$ 总是以 $(x_{(1)} \leqslant x_{(2)} \leqslant \cdots \leqslant x_{(n)})$ 为观测值，则称 $(X_{(1)}, X_{(2)}, \cdots, X_{(n)})$ 为变量 X 的一个顺序统计量。

由顺序统计量出发，可以构造许多有用统计量。例如：

- 样本最大值 $X_{\max} = \max(X_1, X_2, \cdots, X_n) = X_{(n)}$。
- 样本最小值 $X_{\min} = \min(X_1, X_2, \cdots, X_n) = X_{(1)}$。
- 样本极差 $R = \max(X_1, X_2, \cdots, X_n) - \min(X_1, X_2, \cdots, X_n) = X_{(n)} - X_{(1)}$。
- 样本中位数 $\hat{m} = \begin{cases} X_{\frac{n+1}{2}}, & n\text{ 为奇数} \\ \frac{1}{2}\left(X_{\frac{n}{2}} + X_{\frac{n}{2}+1}\right), & n\text{ 为偶数} \end{cases}$

2.6.2 经验分布函数

还可以做出与总体分布函数 $F(x)$ 相应的统计量——经验分布函数，方法如下：设 $X_1, X_2, \cdots, X_n$ 是总体 X 的一个样本，用 $S(x)$ $(-\infty < x < \infty)$ 表示 $X_1, X_2, \cdots, X_n$ 中不大于 x 的随机变量的个数，定义经验分布函数 $F_n(x)$ 为

$$F_n(x) = \frac{1}{n}S(x)\text{，} \quad -\infty < x < \infty \tag{2-30}$$

对于一个样本值，经验分布函数 $F_n(x)$ 的观测值是很容易得到的。$F_n(x)$ 的观测值仍以

$F_n(x)$ 表示。例如：

1）设总体 X 具有一个样本值 1，2，3，则经验分布函数 $F_3(x)$ 的观测值为

$$F_3(x)=\begin{cases}0, & x<1\\ \dfrac{1}{3}, & 1\leqslant x<2\\ \dfrac{2}{3}, & 2\leqslant x<3\\ 1, & x\geqslant 3\end{cases}$$

2）设总体 X 具有一个样本值 1，1，2，则经验分布函数 $F_3(x)$ 的观测值为

$$F_3(x)=\begin{cases}0, & x<1\\ \dfrac{2}{3}, & 1\leqslant x<2\\ 1, & x\geqslant 2\end{cases}$$

一般情况下，设 $x_1,x_2,\cdots,x_n$ 是总体 X 的一个容量为 n 的样本值，先将 $x_1,x_2,\cdots,x_n$ 按自小到大的次序排列，并重新编号，设为

$$x_{(1)}\leqslant x_{(2)}\leqslant\cdots\leqslant x_{(n)}$$

则经验分布 $F_n(x)$ 的观测值为

$$F_n(x)=\begin{cases}0, & x<x_{(1)}\\ \dfrac{k}{n}, & x_{(k)}\leqslant x<x_{(k+1)}\text{，}\quad k=1,2,\cdots,n-1\\ 1, & x\geqslant x_{(n)}\end{cases}$$

对于经验分布函数 $F_n(x)$，格里汶科（Glivenko）在 1933 年证明了以下结果：对于任一实数 x，当 $n\to\infty$ 时 $F_n(x)$ 以概率 1 一致收敛于分布函数 $F(x)$，即

$$P\left\{\min_{n\to\infty}\sup_{-\infty<x<\infty}\left|F_n(x)-F(x)\right|=0\right\}=1 \tag{2-31}$$

由此可见，当 n 充分大时，经验分布函数 $F_n(x)$ 是总体分布函数 $F(x)$ 的一个近似，这就是数量统计中用样本推断总体的理论依据。

【例 2-33】 钢材中的含硅量 X 是影响材料性能的一项重要因素。在炼钢生产的过程中，由于各种随机因素的影响，各炉钢的含硅量 X 是有差异的。对含硅量 X 的概率分布的了解是有关钢材料性能分析的重要依据。某炼钢厂 120 炉正常生产的 25MnSi 钢的含硅量（单位：%）如下：

```
0.86 0.83 0.77 0.81 0.81 0.80 0.79 0.82 0.82 0.81
0.82 0.78 0.80 0.81 0.87 0.81 0.77 0.78 0.77 0.78
0.77 0.71 0.95 0.78 0.81 0.79 0.80 0.77 0.76 0.82
0.84 0.79 0.90 0.82 0.79 0.82 0.79 0.86 0.81 0.78
0.82 0.78 0.73 0.84 0.81 0.81 0.83 0.89 0.78 0.86
0.78 0.84 0.84 0.75 0.81 0.81 0.74 0.78 0.76 0.80
0.75 0.79 0.85 0.78 0.74 0.71 0.88 0.82 0.76 0.85
0.81 0.79 0.77 0.81 0.81 0.87 0.83 0.65 0.64 0.78
```

0.80 0.80 0.77 0.84 0.75 0.83 0.90 0.80 0.85 0.81
0.82 0.84 0.85 0.84 0.82 0.85 0.84 0.82 0.85 0.84
0.81 0.77 0.82 0.83 0.82 0.74 0.73 0.75 0.77 0.78
0.87 0.77 0.80 0.75 0.82 0.78 0.78 0.82 0.78 0.78



求 25MnSi 钢含硅量数据的经验分布函数。

经验分布函数是一种在大样本条件下估计变量分布形态的重要工具。经验分布函数的图像与累积频率折线图在性质上是一致的，它们的主要区别在数据的分组上，经验分布函数处理得更加细腻。

在应用中，可以将经验分布函数图像与可能的分布类型的分布函数图像进行比较，得出关于变量分布形态的结论。

经验分布函数图像的 MATLAB 绘图命令是 cdfplot，其输入参数为样本数据向量，有两个可选输出参数：第一个是图形句柄；第二个是关于样本数据的几个重要的统计量，包括样本最小值、最大值、均值、中值和标准差。

```
>> clear
X=[0.86 0.83 0.77 0.81 0.81 0.80 0.79 0.82 0.82 0.81...
   0.82 0.78 0.80 0.81 0.87 0.81 0.77 0.78 0.77 0.78...
   0.77 0.71 0.95 0.78 0.81 0.79 0.80 0.77 0.76 0.82...
   0.84 0.79 0.90 0.82 0.79 0.82 0.79 0.86 0.81 0.78...
   0.82 0.78 0.73 0.84 0.81 0.81 0.83 0.89 0.78 0.86...
   0.78 0.84 0.84 0.75 0.81 0.81 0.74 0.78 0.76 0.80...
   0.75 0.79 0.85 0.78 0.74 0.71 0.88 0.82 0.76 0.85...
   0.81 0.79 0.77 0.81 0.81 0.87 0.83 0.65 0.64 0.78...
   0.80 0.80 0.77 0.84 0.75 0.83 0.90 0.80 0.85 0.81...
   0.82 0.84 0.85 0.84 0.82 0.85 0.84 0.82 0.85 0.84...
   0.81 0.77 0.82 0.83 0.82 0.74 0.73 0.75 0.77 0.78...
   0.87 0.77 0.80 0.75 0.82 0.78 0.78 0.82 0.78 0.78];
[h,stats]=cdfplot(X)
```

运行程序，输出如下：

```
h =
   172.0016
stats =
         min: 0.6400
         max: 0.9500
        mean: 0.8026
      median: 0.8100
         std: 0.0450
```

由图 2-8 可以看出，样本经验分布函数图像上升速度较快，均值与中值接近，图像的 S 形状均衡对称，均值处函数值约为 0.5。这些特征表明，25MnSi 钢的含硅量可能服从均值为 0.8026、标准差为 0.045 的正态分布。

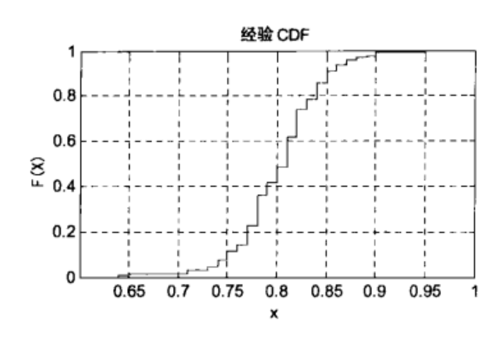


图 2-8　120 炉 25MnSi 钢含硅量数据的经验分布函数图像

2.6.3 χ^2 分布

χ^2 分布分为中心 χ^2 分布和非中心 χ^2 分布两种。

1. 中心 χ^2 分布

中心 χ^2 分布的随机变量由若干独立同分布的零均值高斯变量的平方和得出。设有 n 个独立同分布的零均值高斯随机数 $x_i \sim N(0,\sigma^2)$，$i=1,2,\cdots,n$，则随机数

$$y=\sum_{i=1}^{n} x_i^2 \tag{2-32}$$

服从自由度 n 的中心 χ^2 分布，其概率密度函数为

$$p(y)=\frac{1}{\sigma^2 2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} y^{\frac{n}{2}-1}\exp\left(-\frac{y}{2\sigma^2}\right),\quad y \geqslant 0 \tag{2-33}$$

式中，$\Gamma(x)$ 是伽玛函数，在 MATLAB 中可通过命令 gamma(x)求出，其定义是

$$\Gamma(x)=\int_0^{\infty} \mathrm{e}^{-t} t^{x-1} \mathrm{d}t,\quad x>0 \tag{2-34}$$

特别指出，当 x 为正整数时，有 $\Gamma(x)=(x-1)!$，当 x 为正整数加上 $\frac{1}{2}$ 时，有

$$\Gamma\left(\frac{1}{2}\right)=\sqrt{\pi},\quad \Gamma\left(\frac{3}{2}\right)=\frac{\sqrt{\pi}}{2} \tag{2-35}$$

$$\Gamma\left(m+\frac{1}{2}\right)=\frac{(2m-1)!}{2^m}\sqrt{\pi} \tag{2-36}$$

式中，$(2m-1)!=1\times3\times5\times\cdots\times(2m-1)$，$m=1,2,\cdots$。

自由度为 n 的中心 χ^2 分布随机变量 Y 的期望和方差分别为

$$E(Y)=n\sigma^2 \tag{2-37}$$

$$V(Y)=2n\sigma^4 \tag{2-38}$$

MATLAB 中给出了 $\sigma^2=1$ 的自由度为 n 的中心 χ^2 分布的计算函数：χ^2 分布的分布函数 chi2cdf，分布函数的反函数 chi2inv，概率密度函数 chi2pdf，随机数发生函数 chi2rnd 和期望及方差计算函数 chi2stat 等。



【例 2-34】 设某随机变量服从正态分布，试验得出其 10 个样本为

$$\{1490\ 1440\ 1680\ 1610\ 1500\ 1750\ 1550\ 1420\ 1800\ 1580\}$$

能否认为其期望值 $\mu_0 = 1600$，其方差 $\sigma_0^2 = 14400$（取显著性水平 $\alpha = 0.02$）？

其实现的 MATLAB 程序代码如下：

```
>>clear;
x=[1490 1440 1680 1610 1500 1750 1550 1420 1800 1580];   %样本
m0=1600;                               %给定的期望值
n=length(x);                           %样本数
xbar=mean(x);                          %样本平均
s=std(x,1);                            %样本标准差(有偏)
al=0.02;                               %显著性水平
% 期望的假设检验
t1=tinv(1-al/2,n-1)                    %自由度 n-1 的 t 分布 al/2 分位点
t=(xbar-m0)./(s./sqrt(n-1))            %计算统计量
h_mean=(t>abs(t1))                     %判断:若拒绝,则 h_mean 等于 1
%方差的假设检验
sig2=14400;                            %给定的方差值
ch_1=chi2inv(al/2,n-1)                 %1-al/2 分位点
ch_2=chi2inv(1-al/2,n-1)               %al/2 分位点
ch=n*s^2/(sig2)                        %计算统计量
h_var=(ch<ch_1)|(ch>ch_2)              %判断:若拒绝,则 h_var 等于 1
```

运行程序，输出如下：

```
t1 =
     2.8214        %故均值统计量接受区间为（–2.8214, 2.8214）
t =
    -0.4427        %计算统计量 t 在接受区间内
h_mean =
      0            %接受假设，即有 0.98 的把握说总体期望为 1600
ch_1 =
     2.0879
ch_2 =
    21.6660        %故方差统计量接受区间为（2.0879，21.6660）
ch =
    10.3306        %计算统计量 ch 在接受区间内
h_var =
      0            %接受假设，即有 0.98 的把握说总体方差为 14400
```

2. 非中心 χ^2 分布

非中心 χ^2 分布的随机变量由若干独立同方差的均值不全为零的高斯变量的平方和得出。设有 n 个独立的高斯随机数 $x_i \sim N(m_i,\sigma^2)$，$i=1,2,\cdots,n$，其均值为 m_i，方差均为 σ^2，并设 $s^2 = \sum\limits_{i=1}^{n} m_i^2$，则随机数

$$y=\sum_{i=1}^{n}x_i^2 \tag{2-39}$$

服从自由度为 n 的非中心 χ^2 分布，其概率密度函数为

$$p(y)=\frac{1}{2\sigma^2}\left(\frac{y}{s^2}\right)^{\frac{n-2}{4}}\exp\left(-\frac{s^2+y}{2\sigma^2}\right)I_{\frac{n}{2}-1}\left(\sqrt{y}\frac{s}{\sigma^2}\right),\quad y\geqslant 0 \tag{2-40}$$

式中，$I_a(x)$ 为第一类 a 阶修正贝赛尔函数，MATLAB 提供的计算指令是 besseli(a, x)。自由度为 n 的非中心 χ^2 分布的随机变量 Y 的期望和方差分别为

$$E(Y)=n\sigma^2+s^2 \tag{2-41}$$

$$V(Y)=2n\sigma^4+4\sigma^2s^2 \tag{2-42}$$

MATLAB 统计工具箱给出了指令 ncx2pdf，ncx2cdf，ncx2inv，ncx2rnd，以及 ncx2stat 来计算 $\sigma^2=1$ 的非中心 χ^2 分布问题。

在 MATLAB 命令窗口输入以下代码：

```
>> x=(0:0.1:10)';
p1=ncx2pdf(x,4,2);
p=chi2pdf(x,4);
plot(x,p,'--',x,p1,'-')
```

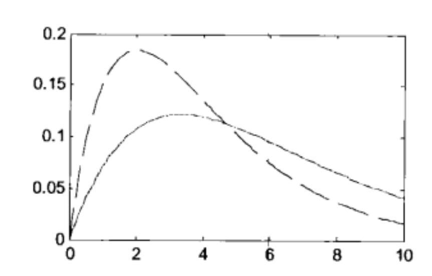
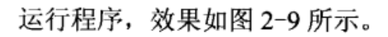


图 2-9　非中心 χ^2 分布的概率密度函数



运行程序，效果如图 2-9 所示。

2.6.4 t 分布

设随机变量 X 和 Y 独立，并且 X 服从正态分布 $N(0,1)$，Y 服从自由度为 n 的 χ^2 分布（概率密度函数为式（2-33），其中 $\sigma=1$），则随机变量

$$t=\frac{X}{\sqrt{Y/n}} \tag{2-43}$$

服从自由度为 n 的 t 分布，其概率密度函数为

$$t=\frac{X}{\sqrt{Y/n}}p_t(x)=\frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi\Gamma\left(\frac{n}{2}\right)}}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}} \tag{2-44}$$

MATLAB 统计工具箱提供了 t 分布的计算指令，包括 tpdf，tcdf，tinv，trnd，tstat 等。

当自由度 $n\to\infty$ 时，t 分布将趋近于标准正态分布。工程上，当 $n>30$ 时，即可将 t 分布视为标准正态分布。

t 分布还可以推广为非中心的 t 分布。MATLAB 统计工具箱也提供了非中心 t 分布的计算指令，包括 nctpdf，nctcdf，nctinv，nctrnd，nctstat 等。

2.6.5 F 分布

设随机变量 X 和 Y 相互独立，分别服从自由度为 m 和 n 的 χ^2 分布（密度函数为

式（2-33），其中$\sigma=1$），即$X\sim\chi^2(m)$，$Y\sim\chi^2(n)$，那么随机变量

$$F=\frac{X/m}{Y/n} \tag{2-45}$$

服从自由度为(m,n)的F分布，其概率密度函数为

$$p_F(x)=\frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)}\left(\frac{m}{n}\right)^{\frac{m}{2}}x^{\frac{m-2}{2}}\left(1+\frac{m}{n}x\right)^{-\frac{m+n}{2}},\quad x\geqslant 0 \tag{2-46}$$

F 分布常用于两个独立χ^2分布随机变量相除运算的问题。显然，一个自由度为(m,n)的F分布随机变量的倒数也服从F分布，但其自由度变为(n,m)。

MATLAB 统计工具箱提供了F分布的计算指令，包括 fpdf，fcdf，finv，frnd，fstat 等。

在 MATLAB 命令窗口中输入以下代码：

```
>> y=fpdf(1:6,2,2)
y =
    0.2500    0.1111    0.0625    0.0400    0.0278    0.0204
>> z=fpdf(3,5:10,5:10)
z =
    0.0689    0.0659    0.0620    0.0577    0.0532    0.0487
```

F 分布还可以推广为非中心的F分布。MATLAB 统计工具箱也提供了非中心F分布的计算指令，包括 ncfpdf，ncfcdf，ncfinv，ncfrnd，ncfstat 等。

在 MATLAB 命令窗口中输入以下代码：

```
>> x=(0.01:0.1:10.01)';
p1=ncfpdf(x,5,20,10);
p=fpdf(x,5,20);
plot(x,p,'-',x,p1,':')
```

运行程序，效果如图 2-10 所示。

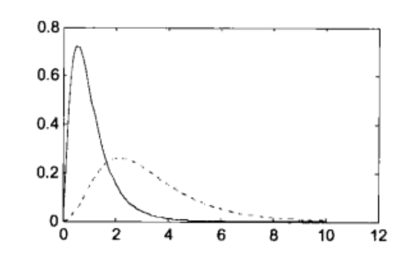


图 2-10　非中心F分布的概率密度函数

2.6.6 超几何分布

设一批产品共M个，其中有K个次品，则任意抽出的$N(N\leqslant M)$个样品中含有的次品数是一个在取值区间$[0,n]$上的离散随机变量。如果用X表示，那么X服从参数为M,K,N的超几何分布。若次品数为x的概率用$P(X=x)=P_x(M,K,N)$表示，则

$$P_x(M,K,N)=\frac{\binom{K}{x}\binom{M-K}{N-x}}{\binom{M}{N}},\quad x=0,1,\cdots,N \tag{2-47}$$

MATLAB 统计工具箱提供了超几何分布的计算指令，包括 hygepdf，hygecdf，hygeinv，hygernd，hygestat 等。

【例 2-35】 如果 100 张软盘，其中 20 张是坏盘，那么抽 10 张出来，坏盘的张数是 0～5 张的概率分别是多少呢？

在 MATLAB 命令窗口输入以下代码：

```
>> p=hygepdf(0:5,100,20,10)
```

运行程序，输出如下：

```
p =
    0.0951    0.2679    0.3182    0.2092    0.0841    0.0215
```

2.6.7 正态分布

设连续型随机变量 X 的概率密度函数为

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}},\quad -\infty<x<+\infty \tag{2-48}$$

式中，μ,σ $(\sigma>0)$ 为常数，则称 X 服从参数为 μ,σ^2 的正态分布或高斯（Gauss）分布，记为 $X\sim N(\mu,\sigma^2)$。

由式（2-48）得 X 的分布函数为

$$F(x)=\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{x}\mathrm{e}^{-\frac{(t-\mu)^2}{2\sigma^2}}\mathrm{d}t \tag{2-49}$$

特别地，当 $\mu=0$，$\sigma=1$ 时，称 X 服从标准正态分布。其概率密度函数和分布函数分别用 $\varphi(x)$，$\Phi(x)$ 表示，即有

$$\varphi(x)=\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{x^2}{2}} \tag{2-50}$$

$$\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}\mathrm{e}^{-\frac{t^2}{2}}\mathrm{d}t \tag{2-51}$$

易知 $\Phi(-x)=1-\Phi(x)$。

$\Phi(x)$ 的函数表可参考附录 A。

一般情况下，若 $X\sim N(\mu,\sigma^2)$，只要通过一个线性变换就能将它化成标准正态分布。

定理 2-8 若 $X\sim N(\mu,\sigma^2)$，则 $Z=\frac{X-\mu}{\sigma}\sim N(0,1)$。

证明： $Z=\frac{X-\mu}{\sigma}$ 的分布函数为

$$P\{Z\leqslant x\}=P\left\{\frac{X-\mu}{\sigma}\leqslant x\right\}=P\{X\leqslant\mu+\sigma x\}=\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\mu+\sigma x}\mathrm{e}^{-\frac{(t-\mu)^2}{2\sigma^2}}\mathrm{d}t$$

令 $\frac{t-\mu}{\sigma}=u$，得

$$P\{Z\leqslant x\}=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}\mathrm{e}^{-\frac{u^2}{2}}\mathrm{d}u=\Phi(x)$$

由此知 $Z=\frac{X-\mu}{\sigma}\sim N(0,1)$。



于是，若 $X \sim N(\mu,\sigma^2)$，则它的分布函数 $F(x)$ 可写成

$$F(x)=P\{X \leqslant x\}=P\left\{\frac{X-\mu}{\sigma} \leqslant \frac{x-\mu}{\sigma}\right\}=\Phi\left(\frac{x-\mu}{\sigma}\right) \tag{2-52}$$

对于任意区间 $(x_1,x_2]$，有

$$P\{x_1 < X \leqslant x_2\}=P\left\{\frac{x_1-\mu}{\sigma} < \frac{X-\mu}{\sigma} \leqslant \frac{x_2-\mu}{\sigma}\right\}=\Phi\left(\frac{x_2-\mu}{\sigma}\right)-\Phi\left(\frac{x_1-\mu}{\sigma}\right) \tag{2-53}$$

为了便于应用，对于标准正态随机变量，引入了 α 分位点的定义。

设 $X \sim N(0,1)$，若 Z_α 满足条件

$$P\{X > Z_\alpha\}=\alpha\text{，}\quad 0<\alpha<1$$

则称点 Z_α 为标准正态分布上的 α 分位点。

【例 2-36】 生成正态分布的随机数。

其实现的 MATLAB 程序代码如下：

```
>> clear;
%设置正态分布的参数
mu0=log(1000);
sigma0=1;
%产生 len 个随机数
len=5;
y1=normrnd(mu0,sigma0,[1 len])
%产生 P 行 Q 列的矩阵
P=3;Q=4;
y2=normrnd(mu0,sigma0,P,Q)
%显示正态分布的柱状图
M=1000;
y3=normrnd(mu0,sigma0,[1,M]);
figure;
t=0:0.1:max(y3);
hist(y3,t);
axis([0 max(y3) 0 50]);
xlabel('取值');ylabel('计数值');
```

运行程序，输出如下：

```
y1 =
    6.4752    5.2422    7.0331    7.1954    5.7613
y2 =
    8.0987    7.2350    7.6335    6.7714
    8.0969    7.0824    6.3194    7.0217
    6.8701    6.7210    9.0909    7.9745
```

正态分布的柱状图如图 2-11 所示。

正态分布在概率论中起着非常重要的作用，在各种分布中，它居于首要地位。在实际中常常遇到一些变量，它们的分布近似于正态分布。例如，在同一生产条件下制造的电灯泡，

使用时数 X 随着每个灯泡而不同。譬如说，第一个可用 1200h，第二个可用 1280h 等，因此 X 总是一个随机变量。实践证明，X 的分布是近似正态的。俗话说的“中间大，两头小”就是正态分布的一个性质。一般来说，在生产条件不变的前提下，许多产品的某些量度（如青砖的抗压强度、细纱的强力、螺钉的口径等）都近似地服从正态分布。这种情况在许多自然科学领域中都存在。例如，热力学中理想气体分子的速度分量、射击时命中位置对目标沿某些标轴的偏差、物理学中测量同一物体的测量误差、生物学中同一生物机体的某一量度（如身长、体重）、某地区一年中的降水量等，都是如此。

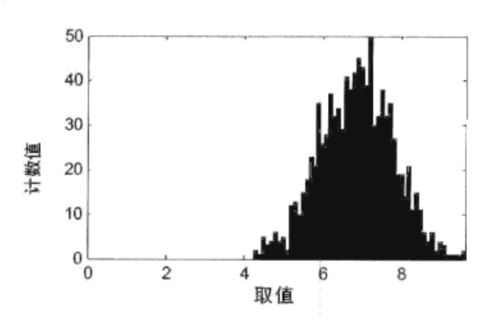


图 2-11　正态分布的柱状图

上述各种量有一共同特点：它们可以被看成许多微小、独立的随机因素的总后果。例如，灯泡的使用时数受原料、工艺、保管条件等因素的影响，而每种因素，在正常情况下，都不能起代替一切的主导作用。具有这种特点的变量一般都可被认为服从正态分布，这一结论的准确的数学叙述见 3.5.3 节的中心极限定理。

2.6.8 正态总体的样本均值与样本方差的分布

确定某个统计量的分布不仅困难，有时甚至是不可能的。现在，我们对总体 X 服从正态分布的情形已经有了详细的研究，下面讲解服从正态分布的总体的统计量的分布。若样本函数 $g(X_1,X_2,\cdots,X_n)$ 含有未知量时，则应称为样本函数的分布。

（1）单个正态总体的统计量的分布

设总体 X（不管服从什么分布，只要均值和方差存在）的均值为 μ，方差为 σ^2，设 $X_1,X_2,\cdots,X_n$ 是来自 X 的一个样本，$\bar{X}$，S^2 是样本均值和样本方差，则总有

$$E(\bar{X})=\mu,\quad D(\bar{X})=\frac{\sigma^2}{n} \tag{2-54}$$

而

$$\begin{aligned}E(S^2)&=E\left[\frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2-n\bar{X}^2\right)\right]=\frac{1}{n-1}\left[\sum_{i=1}^{n}E(X_i^2)-nE(\bar{X}^2)\right]\\&=\frac{1}{n-1}\left[\sum_{i=1}^{n}\left(\sigma^2+\mu^2\right)-n\left(\frac{\sigma^2}{n}+\mu^2\right)\right]=\sigma^2\end{aligned}$$

即

$$E(S^2)=\sigma^2$$

定理 2-9　设 $X_1,X_2,\cdots,X_n$ 是来自正态总体 $N(\mu,\sigma^2)$ 的样本，$\bar{X}$ 是样本均值，则有

$$\bar{X}\sim N\left(\mu,\frac{\sigma^2}{n}\right) \tag{2-55}$$

定理 2-10　设 $X_1,X_2,\cdots,X_n$ 是来自正态总体 $N(\mu,\sigma^2)$ 的样本，$\bar{X}$ 是样本均值，则有

$$Z=\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\sim N(0,1) \tag{2-56}$$

证明：由定理 2-9 知 $\bar{X}\sim N\left(\mu,\frac{\sigma^2}{n}\right)$，所以将 $\bar{X}$ 标准化，即得 $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\sim N(0,1)$。



定理 2-11 设 $X_1,X_2,\cdots,X_n$ 是总体 $N(\mu,\sigma^2)$ 的样本，则有

$$\chi^2=\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2\sim\chi^2(n) \tag{2-57}$$

证明：因为 $X_i\sim N\left(\mu,\sigma^2\right)$，所以

$$\frac{X_i-\mu}{\sigma}\sim N(0,1)\text{，}\quad i=1,2,\cdots,n \tag{2-58}$$

又因为 $X_1,X_2,\cdots,X_n$ 相互独立，所以 $\frac{X_1-\mu}{\sigma},\frac{X_2-\mu}{\sigma},\cdots,\frac{X_n-\mu}{\sigma}$ 也相互独立，于是由 χ^2 分布的定义，可知

$$\chi^2=\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2=\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2\sim\chi^2(n)$$

定理 2-12 设 $X_1,X_2,\cdots,X_n$ 是总体 $N(\mu,\sigma^2)$ 的样本，$\bar{X}$，S^2 分别是样本均值和样本方差，则有

1）$\bar{X}$ 与 S^2 独立。

2）$\chi^2=\frac{(n-1)S^2}{\sigma^2}\sim\chi^2(n-1)$。

这里定理的证明从略，仅对自由度作一些说明：由样本方差 S^2 的定义易知

$$(n-1)S^2=\sum_{i=1}^{n}(X_i-\bar{X})^2$$

所以有

$$\chi^2=\frac{(n-1)S^2}{\sigma^2}=\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\bar{X})^2=\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{\sigma}\right)^2$$

虽然是 n 个随机变量的平方和，但是这些随机变量不是相互独立的，因为它们的和恒等于零。

$$\sum_{i=1}^{n}\frac{X_i-\bar{X}}{\sigma}=\frac{1}{\sigma}\left(\sum_{i=1}^{n}X_i-n\bar{X}\right)\equiv 0$$

由于受到一个条件的约束，所以自由度为 $n-1$。

定理 2-13 设 $X_1,X_2,\cdots,X_n$ 是总体 $N(\mu,\sigma^2)$ 的样本，$\bar{X}$，S^2 分别是样本均值和样本方差，则有

$$t=\frac{\bar{X}-\mu}{S/\sqrt{n}}\sim t(n-1) \tag{2-59}$$

证明：由定理 2-10 知

$$Z=\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\sim N(0,1)$$

因为 $\bar{X}$ 与 S^2 相互独立，所以 $t=\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ 与 $\chi^2=\frac{(n-1)S^2}{\sigma^2}$ 也相互独立，于是由 t 分布的定义知

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\Bigg/\sqrt{\frac{(n-1)S^2}{\sigma^2(n-1)}}\sim t(n-1)$$

化简上式左边，即得

$$\frac{\bar{X}-\mu}{S/\sqrt{n}}\sim t(n-1)$$

（2）两个正态总体的统计量的分布

对于两个正态总体的样本均值和样本方差有以下的定理。设 $X_1,X_2,\cdots,X_{n_1}$ 与 $Y_1,Y_2,\cdots,Y_{n_2}$ 分别是来自正态总体 $N(\mu_1,\sigma_1^2)$ 和 $N(\mu_2,\sigma_2^2)$ 的样本，假设所有的抽样都是相互独立的。由此得到样本 $X_i(i=1,2,\cdots,n_1)$ 与 $Y_j(j=1,2,\cdots,n_2)$ 都是相互独立的随机变量。设 $\bar{X}=\frac{1}{n_1}\sum_{i=1}^{n_1}X_i$，$\bar{Y}=\frac{1}{n_2}\sum_{j=1}^{n_2}Y_j$ 分别是这两个样本的均值；$S_1^2=\frac{1}{n_1-1}\sum_{i=1}^{n_1}(X_i-\bar{X})^2$，$S_2^2=\frac{1}{n_2-1}\sum_{j=1}^{n_2}(Y_j-\bar{Y})^2$ 分别是这两样本的样本方差。

定理 2-14 设总体 X 服从正态分布 $N(\mu_1,\sigma_1^2)$，总体 $Y\sim N(\mu_2,\sigma_2^2)$，则有

$$U=\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$$

服从标准正态分布 $N(0,1)$，即

$$U=\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\sim N(0,1) \tag{2-60}$$

证明：由定理 2-9 知

$$\bar{X}\sim N\left(\mu_1,\frac{\sigma_1^2}{n_1}\right),\quad \bar{Y}\sim N\left(\mu_2,\frac{\sigma_2^2}{n_2}\right)$$

因为 $\bar{X}$ 与 $\bar{Y}$ 相互独立，所以有

$$\bar{X}-\bar{Y}\sim N\left(\mu_1-\mu_2,\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right)$$

于是对 $\bar{X}-\bar{Y}$ 标准化得

$$U=\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\sim N(0,1)$$

特别地，当$\sigma_1=\sigma_2=\sigma$时，可得如下推论。

推理 2-1 设总体X服从正态分布$N(\mu_1,\sigma^2)$，总体Y服从正态分布$N(\mu_2,\sigma^2)$，则有

$$T=\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{S_W\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

服从自由度为n_1+n_2-2的t分布，即

$$T=\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{S_W\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\sim t(n_1+n_2-2) \tag{2-61}$$

其中

$$S_W=\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}} \tag{2-62}$$

证明：由定理 2-14 的推论可知

$$U=\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\sim N(0,1)$$

又由定理 2-12 知

$$\frac{(n_1-1)S_1^2}{\sigma^2}\sim\chi^2(n_1-1)，\quad \frac{(n_2-1)S_2^2}{\sigma^2}\sim\chi^2(n_2-1)$$

因为S_1^2与S_2^2相互独立，所以由χ^2分布的可加性知

$$V=\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{\sigma^2}\sim\chi^2(n_1+n_2-2)$$

因为$\bar{X}$与S_1^2相互独立，$\bar{Y}$与S_2^2，所以U与V也相互独立，于是由t分布的定义可知

$$T=\frac{U}{\sqrt{\frac{V}{n_1+n_2-2}}}=\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{S_W\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\sim t(n_1+n_2-2)$$

定理 2-15 设总体X服从正态分布$N(\mu_1,\sigma_1^2)$，总体Y服从正态分布$N(\mu_2,\sigma_2^2)$，则有

$$F=\frac{\sum_{i=1}^{n_1}(X_i-\mu_1)^2/n_1\sigma_1^2}{\sum_{j=1}^{n_2}(Y_j-\mu_2)^2/n_2\sigma_2^2}$$

服从自由度为(n_1,n_2)的F分布，即

$$F=\frac{\sum_{i=1}^{n_1}(X_i-\mu_1)^2/n_1\sigma_1^2}{\sum_{j=1}^{n_2}(Y_j-\mu_2)^2/n_2\sigma_2^2}\sim F(n_1,n_2) \tag{2-63}$$

证明：由定理 2-11 知

$$\chi_1^2 = \frac{1}{\sigma_1^2}\sum_{i=1}^{n_1}(X_i-\mu_1)^2 \sim \chi^2(n_1)$$

$$\chi_2^2 = \frac{1}{\sigma_2^2}\sum_{j=1}^{n_2}(Y_j-\mu_2)^2 \sim \chi^2(n_2)$$

因为所有的 X_i 与 Y_j 都是相互独立的，所有的 χ_1^2 与 χ_2^2 也相互独立。于是，由 F 分布定义知

$$F = \frac{\chi_1^2/n_1}{\chi_2^2/n_2} = \frac{\sum_{i=1}^{n_1}(X_i-\mu_1)^2/n_1\sigma_1^2}{\sum_{j=1}^{n_2}(Y_j-\mu_2)^2/n_2\sigma_2^2} \sim F(n_1,n_2)$$

定理 2-16 设总体 X 服从正态分布 $N(\mu_1,\sigma_1^2)$，总体 Y 服从正态分布 $N(\mu_2,\sigma_2^2)$，则有 $F=\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$ 服从自由度为 (n_1-1,n_2-1) 的 F 分布，即

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1-1,n_2-1) \tag{2-64}$$

证明：由定理 2-12 知

$$\chi_1^2 = \frac{(n_1-1)S_1^2}{\sigma_1^2} \sim \chi^2(n_1-1)$$

$$\chi_2^2 = \frac{(n_2-1)S_2^2}{\sigma_2^2} \sim \chi^2(n_2-1)$$

因为 S_1^2 与 S_2^2 相互独立，所以 χ_1^2 与 χ_2^2 也相互独立。于是，由 F 分布的定义知

$$F = \frac{\chi_1^2/(n_1-1)}{\chi_2^2/(n_2-1)} = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1-1,n_2-1)$$

【例 2-37】 设 $X_1,X_2,\cdots,X_7$ 为总体 $X\sim N(0,0.5^2)$ 的一个样本，求 $P\left(\sum_{i=1}^{7}X_i^2>4\right)$。

解：因为 $X_i\sim N(0,0.5^2)$，$i=1,2,\cdots,7$

$$\frac{X_i-0}{0.5} \sim N(0,1)$$

由 χ^2 分布的定义知，$\sum_{i=1}^{7}\frac{X_i^2}{(0.5)^2}\sim\chi^2(7)$，所以有

$$P\left\{\sum_{i=1}^{7}X_i^2>4\right\} = P\left\{\sum_{i=1}^{7}\frac{X_i^2}{(0.5)^2}>\frac{4}{(0.5)^2}\right\}$$

查表得，$\chi_{0.025}^2(7)=16.013$，即得所求概率

$$P\left\{\sum_{i=1}^{7}X_i^2>4\right\} = 0.025$$



2.6.9 概率密度函数对比——直方图估计法

数据样本的频率直方图是一种近似求解样本概率密度函数的图解方法，也常用于随机数分布的验证中。

设仿真得出的 n 个样本数据为 $\{x_1, x_2, \cdots, x_n\}$，其样本取值范围为

$$[a,b]=[\min_i x_i, \max_i x_i] \tag{2-65}$$

为了得到样本分布的频率直方图，首先将区间 $[a,b]$ 划分为 m 个等间隔的分组区间，分割点 t_i（i=0,1,…,m）为

$$a=t_0<t_1<\cdots<t_m=b \tag{2-66}$$

分割宽度为

$$\Delta=t_{i+1}-t_i=\frac{b-a}{m},\quad i=0,1,\cdots,m-1 \tag{2-67}$$

然后统计样本数据落入区间 $[t_i, t_{i+1})$ 中的个数 r_i（称为频数），再计算出对应的频率 $f_i=r_i/n$。当样本总数 n 充分大时，频率 f_i 趋近于随机变量 ξ 在该区间的概率，即

$$f_i \approx P(t_i \leqslant \xi < t_{i+1}) \tag{2-68}$$

设随机变量 ξ 的概率密度函数为 $f_\xi(x)$，则有

$$P\{t_i \leqslant \xi < t_{i+1}\}=\int_{t_i}^{t_{i+1}} f_\xi(x)\mathrm{d}x \approx f_\xi(x)\Delta \tag{2-69}$$

所以就可以用样本频率来估计其概率密度函数

$$f_\xi(x)=\frac{f_i}{\Delta}=\frac{r_i}{n\Delta},\quad x\in[t_i, t_{i+1}),\quad i=0,1,\cdots,m-1 \tag{2-70}$$

根据上式作出直方图，与已知分布的概率密度函数对比，即可直接地辨识样本所服从的分布类型。当样本数 $n\to\infty$，$\Delta\to 0$ 时，样本频率直方图将趋近于概率密度函数。

然而，仿真得出的样本数总是有限的，这样在直方图法中如何选择分割区间宽度就显得格外重要了。如果区间选得太宽，直方图将显得粗糙；分割区间过细，则直方图的平滑性不够好。实际应用中可以多选择几种分割宽度，从多种直方图的结果中直观地判断并选取比较平滑而且又比较精细的直方图作为结果，样本数越多，可选择越小的分割区间。在实践中发现，有些情况下，选择直方图分割区间数近似等于样本数据个数的平方根值时得出的直方图较好，即选择

$$m=\lfloor\sqrt{n}\rfloor \tag{2-71}$$

$$\Delta=\frac{b-a}{m} \tag{2-72}$$

MATLAB 中提供了直方图的计算和作图函数 hist。

hist 函数的调用格式如下：

```
[r, xout]=hist(Y, t)
[r, xout]=hist(Y,mbins)
```

其中，[r, xout]=hist(Y, t)：其中 Y 为样本向量；t 是分割区间向量；r 是统计输出的频数；xout 是分割区间向量，等于向量 t；[r, xout]=hist(Y,mbins)：mbins 是分割的区间数。

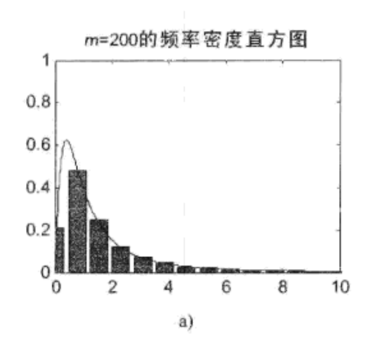
【例 2-38】 试产生自由度为（$n_1=3, n_2=5$）的 F 分布随机数，并用直方图法进行检验。设随机数样本数量为 9999。

其中使用了 frnd 函数来产生 F 分布的随机数，以 hist 函数进行直方图的频数统计，然后转换为频率数据，作出直方图，并修改直方图的样式，最后以 fpdf 函数计算理论概率密度函数并作图比较。

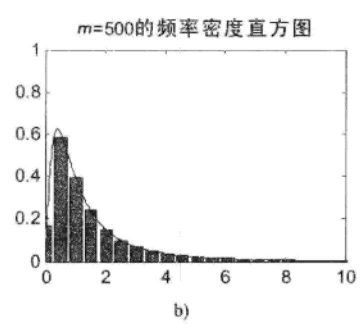
在 M 文件编辑器中输入以下代码。

```
clear;
n1=4;n2=5;                  %F 分布参数
n=10000;                    % 随机数样本数量
x=frnd(n2,n1,n,1);          %随机数样本产生
a=min(x); b=max(x);         %样本值域区间计算
m=200;                      %分组区间数 or m=500 等
de=(b-a)/m;                 %分组宽度
[r,xout]=hist(x,[a:de:b]);  %计算直方图数据
f=r./(n*de);                %计算统计频率密度
bar(xout,f);                %作出频率密度直方图
hold on;
h=findobj(gca,'Type','patch');   %修改直方图样式
set(h,'facecolor',[0.6,0.6,0.6],'edgecolor','k');
x=0:0.01:10;   %计算并画出 F 分布的理论概率密度函数曲线
y=fpdf(x,n2,n1);
plot(x,y,'k-');
axis([0 10 0 1]);
title('m=200 的频率密度直方图');
```

运行程序，效果如图 2-12 所示。图中分别给出了分割区间数为 200 和 500 的直方图结果。



a)



b)

图 2-12 随机分布的直方图检验

a) m=200 的绘图 b) m=500 的绘图

2.7 统计检验

在大多数情况下，分析测试都是采取抽取检验，通过样本测试对总体的某个或某些特征进行估计和作出推断。统计推断包括参数估计与假设检验。参数估计与假设检验是互有联系而又有区别的两类统计推断，参数估计是随机变量分布函数已知，需通过样本估计分布的参数。如果不知道随机变量分布的函数形式，只能假设其具有某种分布形式，假设是否合理，需根据样本值通过检验分布参数来推断其是否正确，属于假设检验。

2.7.1 统计检验的基本原理

统计检验依据的是小概率原理。所谓小概率原理，是指概率很小的事件在一次抽样检验中实际上是不可能发生的。

令检验统计量为

$$T = T(x) \tag{2-73}$$

式中，T 是样本值和被估参数的函数，不包括未知值。在原假设成立时，检验统计量 T 的概率密度函数 $\varphi(T)$ 已知，$\varphi(x)$ 在某个区域 ω 的概率为



$$P(T \in \omega) = \int_{\Gamma \in \omega} \varphi(T) \mathrm{d}T = \alpha \tag{2-74}$$

α 取值通常很小（$\alpha \in$（0.05，0.01）），因此当原假设为真时，T 在区域 ω 内是一个小概率事件。根据小概率原理，在一次抽样检验中几乎是不可能发生的，如果发生了，则有理由认为原假设不正确，这时应在显著性水平 α 下拒绝原假设 H_0，而接受备择假设 H_1。此时，ω 称为拒绝域，拒绝域的边界值称为临界值。

在假设检验时，存在两类错误。第一类错误是当原假设为真时而拒绝原假设，又称为弃真错误。α 为犯假设检验第一类错误的概率，（$1-\alpha$）为原假设为真时作出正确判断的概率。第二类错误就是原假设不成立而错误地接收原假设 H_0，也称为取伪错误。犯第二类错误的概率记为 β。因为统计量是随机变量，因此即使原假设 H_0 不成立，一次抽样检验得到的统计量 T 值也有一定概率落在（$W-\omega$）区域内，则

$$\beta = P(T \in (W-\omega)) = 1 - P(T \in \omega) \tag{2-75}$$

式中，W 为统计量 T 不可能取值的区域。

统计检验的基本方法是概率论反证法，即先成立假设，然后根据小概率原理进行反证。统计检验的一般步骤如下：

1）根据具体问题的要求，建立原假设 H_0 和备择假设 H_1。

2）选择合适的检验统计量。

3）选定显著性水平 α，确定拒绝域 ω。

4）根据样本值计算统计量值。

5）根据小概率原理，使用概率论反证法进行统计推断。若统计量值落在拒绝域内，则拒绝原假设 H_0，而接受备择假设 H_1；若落在非拒绝域内，则接受原假设 H_0；若落在拒绝域与非拒绝域的边界，则怀疑原假设 H_0，此时最好继续进行试验，获得更多的信息，以便作出正确的统计推断。

2.7.2 异常值检验

在一组测定值中，有时会发现一个或几个测定值明显地离群，比其他的测定值明显地偏大或偏小，称为离群值。离群值可能是由实验条件改变、尚不为人所知的新现象突然出现以及系统错误等因素造成的异常值，也可能是由随机误差引起的测定值波动而产生的极值。若为前者，表明离群值与其余的测定值不属于同一总体，应判为异常值；若为后者，尽管极值明显地偏大或偏小，但在统计上仍处于合理的误差限内，与其余测定值属于同一总体，不能将其列为异常值。本节介绍的异常值检验方法都是建立在随机样本测定值服从正态分布和小概率原理基础上的。

异常值的检验可分为两类：一类是标准差已知，另一类是标准差未知。

1．标准已知

（1）两倍和三倍标准差检验法

根据正态分布，出现偏差大于两倍标准差（2σ）和三倍标准差（3σ）的测定值的概率，分别小于 5%和 0.3%是一个小概率事件。如果离群值的偏差大于两倍或三倍标准差，则该离群值应判为异常值。

如果不知道σ，而样本大于 30 时，可直接由样本值计算标准差 s，代替σ来进行检验。

（2）ASTM 检验法

美国材料试验协会提出了一个检验离群值的方法，其检验统计量为

$$T=\frac{|x_d-\bar{x}|}{\sigma} \tag{2-76}$$

式中，x_d 是被检验的异常值，$\bar{x}$ 是一组测定值的平均值，σ 是已知的标准差。若统计量的值大于相应显著性水平α下的临界值T_α，则将 x_d 判为异常值。

（3）NAUR 检验法

若 $x_1\leqslant x_2\leqslant\cdots\leqslant x_n$ 为按大小排列的一个样本值，它服从$N(\mu,\sigma^2)$，则检验统计量为

$$R_n=\frac{|x_d-\bar{x}|}{\sigma} \tag{2-77}$$

当计算的 R_n 值大于相应显著性水平α下的临界值$R_{\alpha,n}$，则将 x_d 判为异常值。

2. 标准差未知

（1）t 检验法

将可疑测定值 x_d 以外的其余测定值当做一个总体，并假设该总体服从正态分布。由这些测定值计算平均值$\bar{x}$与标准差 s，而将可疑值 x_d 当做一个样本容量为 1 的特殊总体。如果 x_d 与其余测定值同属于一个总体，则它与其余测定值之间不应有显著性差异。检测统计量为

$$T=\frac{|x_d-\bar{x}|}{\sigma} \tag{2-78}$$

若统计量的值大于相应显著性水平α下的 t 检验法的临界值T_α，则将 x_d 判为异常值。

（2）极差检验法

可用极差 $R=x_{\max}-x_{\min}$ 来估计标准差 s，因此可用极差来进行异常值检验，统计量为

$$t_R=\frac{|x_d-\bar{x}|}{R} \tag{2-79}$$

若统计量的值大于相应显著性水平α下的极差检验法的临界值$t_{\alpha,R}$，则将 x_d 判为异常值。

不同检验方法的检验功效不同，适用的场合也不同。一个离群值是否判为异常值，与统计检验时采用的检验标准有关。

2.7.3 方差检验

当对一试样进行多次重复测定时，由于受到各种因素的影响，测定值产生随机波动，随机波动的大小，反映了测试条件的稳定性和测定结果的精确度。可用方差来度量。方差检验的目的就是要从统计上检验与判断各方差之间是否存在显著性差异，从分析测试的角度来看，也就是要判断各分析方法或分析结果的精密度是否一致。方差检验在工业生产过程控



制、保证产品质量、保证数据的可比性等方面具有重要的作用。

1．一个总体方差的检验

若 $x_1,x_2,\cdots,x_n$ 为服从正态分布 $N(\mu,\sigma^2)$ 的一个样本值，则 $\sum_{i=1}^{n}(x_i-\mu)/\sigma^2=(n-1)s^2/\sigma^2$ 为服从自由度为 $f=n-1$ 的 χ^2 分布，总体方差 σ^2 的置信区间为

$$\frac{(n-1)s^2}{\chi^2_{a/2}}<\sigma^2<\frac{(n-1)s^2}{\chi^2_{1-a/2}} \tag{2-80}$$

$$P\left(\chi^2_{1-a/2}<\frac{(n-1)s^2}{\sigma^2}<\chi^2_{a/2}\right)=1-a \tag{2-81}$$

当只有一个总体，且总体方差 σ^2 已知时，可用式（2-81）来检验总体方差。统计检验的假设为

$$\begin{aligned}H_0:\sigma^2=\sigma_0^2\\H_1:\sigma^2\neq\sigma_0^2\end{aligned} \tag{2-82}$$

2．两个总体方差的检验

若 $x_1,x_2,\cdots,x_n$ 为服从正态分布 $N(\mu_1,\sigma_1^2)$ 的一个样本值， $y_1,y_2,\cdots,y_n$ 为服从正态分布 $N(\mu_2,\sigma_2^2)$ 的一个样本值，则有

$$F=\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}\sim F(n_1-1,n_2-1)$$

统计检验的假设为

$$\begin{aligned}H_0:\sigma_1^2=\sigma_2^2\\H_1:\sigma_1^2\neq\sigma_2^2\end{aligned} \tag{2-83}$$

当原假设 H_0 真时， $F=S_1^2/S_2^2$ 可用来检验总体方差的齐性。式中，S_1 是两个方差中较大的一个，S_2 是较小的一个。统计量 $F>F_a$，而落在拒绝域的概率为 α 。

3．多个总体方差的检验

假设有 m 个总体，分别服从正态分布 $N(\mu_1,\sigma_1^2),\cdots,N(\mu_m,\sigma_m^2)$，由 m 个总体中分别独立地抽取容量为 $n_1,n_2,\cdots,n_m$ 的样本，各样本的方差为 $s_1^2,s_2^2,\cdots,s_m^2$。现在要检验

$$H_0:\sigma_1^2=\sigma_2^2\cdots=\sigma_m^2 \tag{2-84}$$

常用的方法有 Bartlett 检验法、Cochran 检验法、Hartley 检验法和极差比较检验法。

2.7.4 分布拟合检验

在实际工作中，有时并不知道总体服从什么分布，这就需要根据样本数据来检验总体的分布形式，称为分布拟合检验，其中最常见的是总体分布正态性检验。常用的方法有正态概率纸法、χ^2 检验法等。

【例 2-39】 从一批滚珠中随机抽取 50 个，测得它们的直径（单位：mm）为

15.0,15.8,15.2,15.1,15.9,14.7,14.8,15.5,15.6,15.3,

15.1,15.3,15.0,15.6,15.7,14.8,14.5,14.2,14.9,14.9,
15.2,15.0,15.3,15.6,15.1,14.9,14.2,14.6,15.8,15.2,
15.9,15.2,15.0,14.9,14.8,14.5,15.1,15.5,15.5,15.1,
15.1,15.0,15.3,14.7,14.5,15.5,15.0,14.7,14.6,14.2

是否可以认为这批滚珠的直径服从正态分布呢（$\alpha=0.05$）？求出总体的均值。

分析：该问题可归结为正态分布拟合的检验问题，且样本较大，选用命令 jbtest()，显著性水平 $\alpha=0.05$。

其实现的 MATLAB 程序代码如下：

```
clear;
X=[15.0,15.8,15.2,15.1,15.9,14.7,14.8,15.5,15.6,15.3,...
    15.1,15.3,15.0,15.6,15.7,14.8,14.5,14.2,14.9,14.9,...
    15.2,15.0,15.3,15.6,15.1,14.9,14.2,14.6,15.8,15.2,...
    15.9,15.2,15.0,14.9,14.8,14.5,15.1,15.5,15.5,15.1,...
    15.1,15.0,15.3,14.7,14.5,15.5,15.0,14.7,14.6,14.2];
[h,P,Jbstat,CV]=jbtest(X,0.05)
mu=mean(X)
```

运行程序，输出如下：

```
h =        0
P =      0.5000
Jbstat =      0.4573
CV =      4.9697
mu =     15.0780
```

h=0 表示在显著性水平 $\alpha=0.05$ 下接受原假设，且 P = 0.5000 表明接受假设的概率也很大，测试值 *Jbstat* = 0.4573 小于临界值 *CV* = 4.9697，所以接受原假设。此时，均值为 *mu* = 15.0780。

【例 2-40】 淮河流域历史上经常发生洪水灾害，据统计 1949～1991 年淮河流域成灾面积（单位：万亩）每年总计分别为

3383.4 4687.4 1631.1 2244.5 2011.7 6123.1 1918.0 6232.4
5453.9 1412.4 321.5 2185.0 1285.4 4079.6 10124.2 5532.7
3809.3 389.4 412.1 809.7 870.6 1055.7 1451.8 1532.9 765.9
1987.6 2765.5 739.9 515.6 428.4 3794.5 242.3 4812 2204.7
4407.1 2885 1124.7 1190 191.4 2227.9 2079 6934.1

试检验全流域的成灾面积是否服从正态分布？

分析：该问题可归结为正态分布拟合的检验问题，分别选用概率纸法与命令 jbtest 检验。

其实现的 MATLAB 程序代码如下：

```
>> clear;
X=[3383.4 4687.4 1631.1 2244.5 2011.7 6123.1 1918.0 6232.4...
    5453.9 1412.4 321.5 2185.0 1285.4 4079.6 10124.2 5532.7...
    3809.3 389.4 412.1 809.7 870.6 1055.7 1451.8 1532.9 765.9...
```

```
    1987.6 2765.5 739.9 515.6 428.4 3794.5 242.3 4812 2204.7...
    4407.1 2885 1124.7 1190 191.4 2227.9 2079 6934.1];          %输入原始数据
normplot(X);   %用概率纸检验数据是否服从正态分布
[h,P,Jbstat,CV]=jbtest(X,0.05)   %正态分布拟合的检验
```



运行程序，输出如下（见图 2-13）：

```
h =        1
P =      0.0051
Jbstat =     16.7897
CV =      4.7992
```

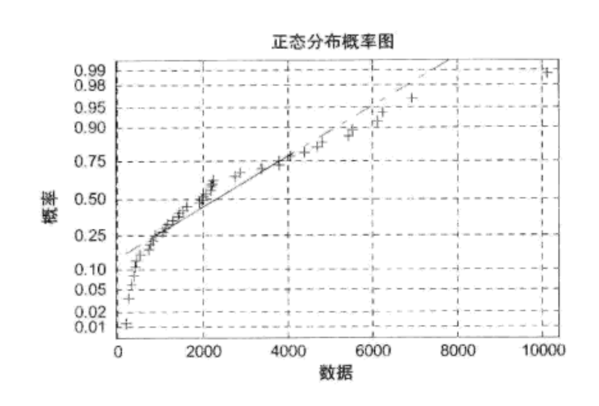


图 2-13 概率纸检验图

从图上可以看出散点并不聚集在直线上，因此流域成灾面积（原始数据）不服从正态分布，这一点也可以通过 jbtest 检验来证实。由于 h=1 表示在置信水平 $\alpha = 0.05$ 下不接受原假设，且 $P = 0.0051$ 表明接受假设的概率也很小，测试值 $Jbstat = 16.7897$ 大于临界值 $CV = 4.7992$，所以不接受原假设。

【例 2-41】 已知数据

```
x=[2 3 4 5 7 8 11 14 15 16 18 19];
y=[106.42 108.2 109.58 110 109.93 110.49 110.59 110.6 110.9 110.76 111 111.2]
```

建立 y 与 x 之间的函数关系，并检验残差 r 是否服从均值为零的正态分布。

分析：通过作散点图，猜测曲线的参数表达式，求出最佳参数，得到 y 与 x 之间的函数关系，计算出残差，检验残差 e_i 是否服从均值为零的正态分布。

其实现的 MATLAB 程序代码如下：

```
>> clear
x=[2 3 4 5 7 8 11 14 15 16 18 19];
y=[106.42 108.2 109.58 110 109.93 110.49 110.59 110.6 110.9 110.76 111 111.2];
plot(x,y,'*');                        %作散点图
A=polyfit(x,y,1)                      %线性最小二乘拟合
plot(x,y,'*',x,polyval(A,x),'r');     %绘制拟合直线
e1=y-polyval(A,x);                    %计算出残差
```

```
[h1,sig,ci]=ttest(e1,0,0.05)              %用 t 检验来检验残差是否服从正态分布
[h2,P,Jbstat,CV]=lillietest(e1,0.05)      %正态分布拟合的检验
```

运行程序，输出如下：

```
A =
    0.1804    108.1387
h1 =      0
sig =     1.0000
ci =
   -0.5307      0.5307
h2 =      0
P =      0.2039
Jbstat =      0.1995
CV =      0.2418
```

x 与 y 的线性最小二乘拟合直线方程为 y=0.1804x+108.1387，不管是 t 检验还是 lillietest()检验，都接受残差 e_i 服从均值为零的正态分布的假设，但要注意函数 lillitest 检验给出的 P = 0.2039 很小，说明虽然通过检验，但不是很理想，这点从拟合的直线（见图 2-14）也能直观地看出来。

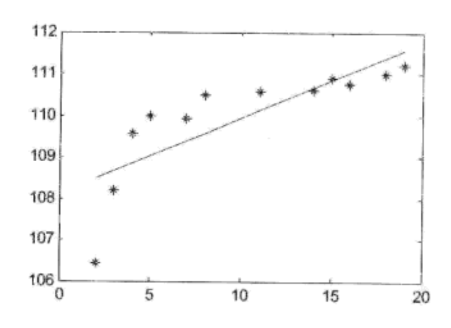


图 2-14　拟合的直线效果



第 3 章　多维随机变量

前面只讲解了一个随机变量的情况，但是在实际问题中，常常需要同时用两个或两个以上的随机变量才能较好地表示一个试验的结果，而这些随机变量之间往往存在一定的联系，因而需要将其作为一个整体来研究。为此，引进随机向量的概念。

3.1　二维随机变量

3.1.1 二维随机变量的定义

在实际问题中，对于某些随机试验的结果往往需要同时用两个随机变量来描述。

例如，为了研究某一地区儿童的身体发育情况，对这一地区的儿童进行抽查，对每个儿童都要观察其身高 X，体重 Y。在这里，X，Y 是两个随机变量。

又如，大炮射击时，炮弹的弹着点的位置需要由其横坐标 X 和纵坐标 Y 来确定。在这里，X，Y 也是两个随机变量。

这类例子很多，举不胜举。值得强调的是，这些例子中的两个随机变量间一般来说都有着某种联系，因此，需要将这两个随机变量作为一个整体来研究。其在数学上的抽象就是如下定义的二维随机变量。

定义 3-1　设 X，Y 是两个随机变量，由它们构成的整体 ξ 称为二维随机向量，记为 (X,Y)，即

$$\xi=(X,Y)$$

3.1.2 离散型随机向量

与随机变量的情形类似，对于二维随机向量，也只讲解离散型和连续型，且首先讲解离散型随机向量。

定义 3-2　如果二维随机向量 $\xi=(X,Y)$ 全部可能取到的值（二维向量）能够一一列举出来，则称 ξ 为二维离散型随机向量。

显然，如果 $\xi=(X,Y)$ 是二维离散型随机向量，则 X，Y 都是离散型随机变量。反之，也成立。

设二维离散型随机向量 $\xi=(X,Y)$ 所有可能取到的值为

$$(x_i,y_i)，\quad i,j=1,2,\cdots$$

且取得各个值的概率为

$$p_{ij}=P[(X,Y)=(x_i,y_i)]，\quad i,j=1,2,\cdots \tag{3-1}$$

则称式（3-1）为 $\xi=(X,Y)$ 的联合概率分布，简称为概率分布或联合分布。

显然，p_{ij} 具有下列性质：

1）$p_{ij} \geqslant 0$，$i,j=1,2,\cdots$

2）$\sum_i \sum_j p_{ij} = 1$。

【例 3-1】 整数 X 随机地在 1，2，3，4 中取一个值，另一个整数 Y 随机地在 1～X 中取一个值，求 (X,Y) 的概率分布。

解： 由概率的乘法公式，可得 (X,Y) 的概率分布为

$$P[(X,Y)=(1,1)]=\frac{1}{4}\times 1=\frac{1}{4}$$

$$P[(X,Y)=(1,2)]=\frac{1}{4}\times 0=0$$

$$P[(X,Y)=(1,3)]=\frac{1}{4}\times 0=0$$

$$P[(X,Y)=(1,4)]=\frac{1}{4}\times 0=0$$

$$P[(X,Y)=(2,1)]=\frac{1}{4}\times \frac{1}{2}=\frac{1}{8}$$

$$P[(X,Y)=(2,2)]=\frac{1}{4}\times \frac{1}{2}=\frac{1}{8}$$

$$P[(X,Y)=(2,3)]=\frac{1}{4}\times 0=0$$

$$P[(X,Y)=(2,4)]=\frac{1}{4}\times 0=0$$

$$P[(X,Y)=(3,1)]=\frac{1}{4}\times \frac{1}{3}=\frac{1}{12}$$

$$P[(X,Y)=(3,2)]=\frac{1}{4}\times \frac{1}{3}=\frac{1}{12}$$

$$P[(X,Y)=(3,3)]=\frac{1}{4}\times \frac{1}{3}=\frac{1}{12}$$

$$P[(X,Y)=(3,4)]=\frac{1}{4}\times 0=0$$

$$P[(X,Y)=(4,1)]=\frac{1}{4}\times \frac{1}{4}=\frac{1}{16}$$

$$P[(X,Y)=(4,2)]=\frac{1}{4}\times \frac{1}{4}=\frac{1}{16}$$

$$P[(X,Y)=(4,3)]=\frac{1}{4}\times \frac{1}{4}=\frac{1}{16}$$

$$P[(X,Y)=(4,4)]=\frac{1}{4}\times \frac{1}{4}=\frac{1}{16}$$

【例 3-2】 箱子中有 12 个产品，其中有两个次品，现从箱子中无放回地抽取两次，每次只取一个产品，且按如下方式定义随机变量 X,Y：

$$X=\begin{cases}0, & \text{第一次取出正品}\\ 1, & \text{第一次取出次品}\end{cases},\quad Y=\begin{cases}0, & \text{第二次取出正品}\\ 1, & \text{第二次取出次品}\end{cases}$$

求随机向量 (X,Y) 的概率分布。

解：由概率的乘法公式，可得随机向量(X,Y)的概率分布为

$$P[(X,Y)=(0,0)]=\frac{10}{12}\times\frac{9}{11}=\frac{90}{132}=\frac{15}{22}$$

$$P[(X,Y)=(0,1)]=\frac{10}{12}\times\frac{2}{11}=\frac{20}{132}=\frac{5}{33}$$

$$P[(X,Y)=(1,0)]=\frac{2}{12}\times\frac{10}{11}=\frac{20}{132}=\frac{5}{33}$$

$$P[(X,Y)=(1,1)]=\frac{2}{12}\times\frac{1}{11}=\frac{2}{132}=\frac{1}{66}$$



3.1.3 连续型随机向量

定义 3-3 对于二维随机向量$\xi=(X,Y)$，如果存在非负可积函数$p(x,y)$，$-\infty<x<\infty,-\infty<y<\infty$，使得对于任意一个由不等式$a<x<b$，$c<y<d$确定的平面区域

$$D=\{(x,y)\,|\,a<x<b,c<y<d\}$$

均有

$$P[(x,y)\in D]=\iint_D p(x,y)\mathrm{d}x\mathrm{d}y \tag{3-2}$$

则称$\xi=(X,Y)$为二维连续型随机向量，且称$p(x,y)$为$\xi=(X,Y)$的联合分布密度，简称为分布密度或（概率）密度。

显然，如果$\xi=(X,Y)$是二维连续型随机向量，则X,Y都是连续型随机变量。反之，也成立。

根据定义 3-3 可知，二维连续型随机向量(X,Y)的分布密度$p(x,y)$具有下列性质：

1）$p(x,y)\geqslant 0$。

2）$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}p(x,y)\mathrm{d}x\mathrm{d}y=1$。

3）对于任意的平面区域D，均有

$$P[(x,y)\in D]=\iint_D p(x,y)\mathrm{d}x\mathrm{d}y$$

上述性质的证明从略。

【例 3-3】 设随机向量(X,Y)的分布密度为

$$p(x,y)=\begin{cases}Ae^{-(2x+7y)}, & x>0,y>0\\ 0, & 其他\end{cases}$$

求常数A。

解：由于

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}p(x,y)\mathrm{d}x\mathrm{d}y=1$$

而

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}p(x,y)\mathrm{d}x\mathrm{d}y=\int_0^{+\infty}\int_0^{+\infty}Ae^{-(2x+7y)}\mathrm{d}x\mathrm{d}y$$

$$=A\int_0^{+\infty}e^{-2x}dx\int_0^{+\infty}e^{-7y}dy=A\left[-\frac{1}{2}e^{-2x}\right]_0^{+\infty}\cdot\left[-\frac{1}{7}e^{-7y}\right]_0^{+\infty}$$
$$=\frac{1}{14}A$$

因此$\frac{1}{14}A=1$，即$A=14$。

3.1.4 随机向量的均匀分布

定义 3-4 设 D 是平面上的有界区域，其面积为 A。如果二维随机向量(X,Y)的分布密度为

$$p(x,y)=\begin{cases}\frac{1}{A}, & (x,y)\in D\\ 0, & 其他\end{cases}$$

则称二维随机向量(X,Y)在区域D上服从均匀分布。

下面主要介绍 MATLAB 中与均匀分布有关的一些函数。

（1）unifpdf 函数

功能：用于计算均匀分布$U(a,b)$的密度函数。

其调用格式如下：

```
Y = unifpdf(x,a,b)
```

其中，$a<b$ 为该分布的参数，而 x 是数或矩阵。此时，函数的计算结果是一个与 x 同维数的矩阵，其各个元素是 x 相应元素的均匀分布$U(a,b)$的分布密度函数值。

其公式为

$$y=f(x\,|\,a,b)=\frac{1}{b-a}I_{[a,b]}(x)$$

例如，其实现的 MATLAB 程序代码如下：

```
>> unifpdf(1:10,0.25,5)
```

运行程序，输出如下：

```
ans =
    0.2105    0.2105    0.2105    0.2105    0.2105    0    0    0    0    0
```

计算结果是一个 10 维的行向量，它的第 k 个分量恰好等于$U(0.25,5)$的密度函数在 k 点的值。

（2）unifcdf 函数

功能：用于计算均匀分布$U(a,b)$的累积分布函数值。该函数的调用格式如下：

```
P=unifcdf(x, a,b)
```

其中，$a<b$ 为该分布的参数，而 x 是数或矩阵。此时，函数的计算结果是一个与 x 同维数的矩阵，其各个元素是 x 相应元素的均匀分布$U(a,b)$的累积分布函数值。

其公式为

$$p = F(x \mid a,b) = \frac{x-a}{b-a} I_{[a,b]}(x)$$

例如，其实现的 MATLAB 程序代码如下：

```
>> probability = unifcdf(0.75,-1,1)
```



运行程序，输出如下：

```
probability =
    0.8750
```

（3）unifrnd 函数

功能：用于生成服从均匀分布 $U(a,b)$ 的随机数，它有 3 种调用格式：

- 第一种调用格式如下：

```
r=unifrnd(a,b)
```

其中，$a<b$ 为该分布的参数。这种调用的计算结果为一个服从均匀分布 $U(a,b)$ 的随机数。

- 第二种调用格式如下：

```
r=unifrnd(a,b, n)
```

其中，$a<b$ 为该分布的参数，n 为以正整数为分量的二维行向量。这种调用的计算结果是一个由服从均匀分布 $U(a,b)$ 的随机数所组成的矩阵，该矩阵的行数由 n 的第一个分量指定，列数由 n 的第二个分量指定。

- 第三种调用格式如下：

```
r=unifrnd(a,b, n, m)
```

其中，$a<b$ 为该分布的参数，n 和 m 为正整数。这种调用的计算结果为一个 $n\times m$ 阶矩阵，其各个元素都是服从均匀分布 $U(a,b)$ 的随机数。

例如，unifrnd 函数实现的 MATLAB 程序代码如下：

```
>> random = unifrnd(0,1:6)
random =
    0.8147    1.8116    0.3810    3.6535    3.1618    0.5852
>> random = unifrnd(0,1:6,[1 6])
random =
    0.2785    1.0938    2.8725    3.8596    0.7881    5.8236
>> random = unifrnd(0,1,2,3)
random =
    0.9572    0.8003    0.4218
    0.4854    0.1419    0.9157
```

3.2 随机向量的分布

3.2.1 边缘分布

二维随机变量(X,Y)作为一个整体，具有联合分布函数$F(x,y)$，而X,Y各自都是随机变量，它们也有自己的分布函数$F_X(x)$，$F_Y(y)$。相对于二维随机变量(X,Y)的联合分布函数，分别称$F_X(x)$，$F_Y(y)$为X和Y的边缘分布函数。相应地，离散型随机变量X,Y各自的分布律称为边缘分布律，而连续型随机变量X,Y各自的概率密度称为边缘概率密度。将边缘分布函数、边缘分布律和边缘概率密度统称为边缘分布。

1．边缘分布函数

设(X,Y)是二维随机变量，其联合分布函数为$F(x,y)$，则X的边缘分布函数为

$$F_X(x)=P\{X\leqslant x\}=P\{X\leqslant x,Y<+\infty\}$$

即

$$F_X(x)=F(x,+\infty)=\lim_{y\to+\infty}F(x,y) \tag{3-3}$$

同理，Y的边缘分布函数为

$$F_Y(y)=F(+\infty,y)=\lim_{x\to+\infty}F(x,y) \tag{3-4}$$

【例 3-4】 设二维连续型随机变量(X,Y)的联合分布函数为

$$F(x,y)=\begin{cases}0, & x<0\text{或}y<0\\ \dfrac{1}{3}x^2y\left(x+\dfrac{y}{4}\right), & 0\leqslant x\leqslant 1,0\leqslant y\leqslant 2\\ \dfrac{1}{3}x^2(2x+1), & 0\leqslant x\leqslant 1,y>2\\ \dfrac{1}{12}y(4+y), & x>1,0\leqslant y\leqslant 2\\ 0, & x>1,y>2\end{cases}$$

求：1）X,Y的边缘分布函数$F_X(x)$，$F_Y(y)$。2）(X,Y)的联合概率密度$f(x,y)$。

解：1）

$$F_X(x)=F(x,+\infty)=\begin{cases}0, & x<0\\ \dfrac{1}{3}x^2(2x+1), & 0\leqslant x\leqslant 1\\ 1, & x>1\end{cases}$$

$$F_Y(y)=F(+\infty,y)=\begin{cases}0, & y<0\\ \dfrac{1}{12}y(4+y), & 0\leqslant y\leqslant 2\\ 1, & y>2\end{cases}$$

2）因为$f(x,y)=\dfrac{\partial^2F(x,y)}{\partial x\partial y}$，而

$$\frac{\partial F(x,y)}{\partial x}=\begin{cases}x^2y+\frac{1}{6}xy^2, & 0\leqslant x\leqslant 1,0\leqslant y\leqslant 2\\ 2x^2+\frac{2}{3}x, & 0\leqslant x\leqslant 1, y>2\\ 0, & 其他\end{cases}$$



$$\frac{\partial^2 F(x,y)}{\partial x\partial y}=\begin{cases}x^2+\frac{1}{3}xy, & 0\leqslant x\leqslant 1,0\leqslant y\leqslant 2\\ 0, & 其他\end{cases}$$

所以(X,Y)的联合概率密度为

$$f(x,y)=\begin{cases}x^2+\frac{1}{3}xy, & 0\leqslant x\leqslant 1,0\leqslant y\leqslant 2\\ 0, & 其他\end{cases}$$

2. 边缘分布律

设(X,Y)为二维离散型随机变量，联合分布律为

$$P\{X=x_i,Y=y_j\}=p_{ij},\quad i,j=1,2,\cdots$$

则

$$P\{X=x_i\}=P\{(X=x_i)\bigcup_{j=1}^{\infty}(Y=y_j)\}=P\{\bigcup_{j=1}^{\infty}(X=x_i,Y=y_j)\}$$

$$=\sum_{j=1}^{\infty}P\{X=x_i,Y=y_j\}=\sum_{j=1}^{\infty}p_{ij}=p_{i\cdot},i=1,2,\cdots$$

同理

$$P\{Y=y_j\}=\sum_{i=1}^{\infty}p_{ij}=p_{\cdot j},j=1,2,\cdots$$

称

$$P\{X=x_i\}=p_{i\cdot},i=1,2,\cdots$$

为X的边缘分布律。

称

$$P\{Y=y_j\}=p_{\cdot j},j=1,2,\cdots$$

为Y的边缘分布律。

若(X,Y)的联合分布律用表格表示，则$p_{i\cdot}$就是表格上第 i 行的元素之和，$p_{\cdot j}$就是表格第j列的元素之和。分别将它们记在表格的边上，如下表。这也是边缘分布名称的由来。

X \ Y	y_1	y_2	$\cdots$	y_j	$\cdots$	$P\{X=x_i\}$
x_1	p_{11}	p_{12}	$\cdots$	p_{1j}	$\cdots$	$p_{1\cdot}$
x_2	p_{21}	p_{22}	$\cdots$	p_{2j}	$\cdots$	$p_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
x_i	p_{i1}	p_{i2}	$\cdots$	p_{ij}	$\cdots$	$p_{i\cdot}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
$P\{Y=y_j\}$	$p_{\cdot 1}$	$p_{\cdot 2}$	$\cdots$	$p_{\cdot j}$	$\cdots$	

【例 3-5】 设有 10 件产品，其中有两件次品，8 件正品。现从中抽取两次，每次取一件产品。定义随机变量 X,Y 如下：

$$X=\begin{cases}1, & \text{第一次取出次品}\\0, & \text{第一次取出正品}\end{cases},\quad Y=\begin{cases}1, & \text{第二次取出次品}\\0, & \text{第二次取出正品}\end{cases}$$

试就下列两种情况，分别求 (X,Y) 的联合分布律和边缘分布律。

1）有放回抽取。2）不放回抽取。

解：1）有放回抽取。

此时，事件 $\{X=x_i\}$ 与 $\{Y=y_j\}$ 相互独立，故

$$P\{X=x_i,Y=y_j\}=P\{X=x_i\}P\{Y=y_j\},i,j=1,2,\cdots$$

即

$$P\{X=0,Y=0\}=P\{X=0\}P(Y=0)=\frac{8}{10}\times\frac{8}{10}=\frac{16}{25}$$

同理

$$P\{X=0,Y=1\}=\frac{8}{10}\times\frac{2}{10}=\frac{4}{25}$$

$$P\{X=1,Y=0\}=\frac{2}{10}\times\frac{8}{10}=\frac{4}{25}$$

$$P\{X=1,Y=1\}=\frac{2}{10}\times\frac{2}{10}=\frac{1}{25}$$

至此，求出 (X,Y) 的联合分布律。

$$P\{X=0\}=P\{X=0,Y=0\}+P\{X=0,Y=1\}=\frac{16}{25}+\frac{4}{25}=\frac{4}{5}$$

$$P\{X=1\}=1-P\{X=0\}=\frac{1}{5}$$

即 X 的边缘分布律为

X	0	1
p_k	$\frac{4}{5}$	$\frac{1}{5}$

同理，可求出 Y 的边缘分布律为

Y	0	1
p_k	$\frac{4}{5}$	$\frac{1}{5}$

(X,Y) 的联合分布律及边缘分布律也可用表格表示，如下所示。

有放回抽取

X \ Y	0	1	$P\{X=x_i\}$
0	$\frac{16}{25}$	$\frac{4}{25}$	$\frac{4}{5}$
1	$\frac{4}{25}$	$\frac{1}{25}$	$\frac{1}{5}$
$P\{Y=y_j\}$	$\frac{4}{5}$	$\frac{1}{5}$	1



2）不放回抽取。

此时，事件$\{X=x_i\}$与$\{Y=y_j\}$不相互独立，由乘法公式知

$$P\{X=i,Y=j\}=P\{X=i\}P\{Y=j\mid X=i\},i,j=0,1$$

即

$$P\{X=0,Y=0\}=P\{X=0\}P(Y=0\mid X=0)=\frac{8}{10}\times\frac{7}{9}=\frac{28}{45}$$

同理

$$P\{X=0,Y=1\}=\frac{8}{10}\times\frac{2}{9}=\frac{8}{45}$$

$$P\{X=1,Y=0\}=\frac{2}{10}\times\frac{8}{9}=\frac{8}{45}$$

$$P\{X=1,Y=1\}=\frac{2}{10}\times\frac{1}{9}=\frac{1}{45}$$

(X,Y)的联合分布律也可用表格表示。按行、列求和，即得X和Y的边缘分布律。可用如下表格表示：

不放回抽取

X \ Y	0	1	$P\{X=x_i\}$
0	$\frac{28}{45}$	$\frac{8}{45}$	$\frac{4}{5}$
1	$\frac{8}{45}$	$\frac{1}{45}$	$\frac{1}{5}$
$P\{Y=y_j\}$	$\frac{4}{5}$	$\frac{1}{5}$	1

从以上两个表格看到，在两种不同情形下，X,Y的边缘分布律相同，但(X,Y)的联合分布律不同。因此可以看出，仅由X和Y的边缘分布不能确定(X,Y)的联合分布。

3．边缘概率密度

设(X,Y)为二维连续型随机变量，联合概率密度为$p(x,y)$，而X,Y各自的概率密度$p_X(x)$，$p_Y(y)$为边缘概率密度，它们由联合概率密度决定。

由式（3-2）和式（3-3）可得

$$F_X(x)=F(x,+\infty)=\int_{-\infty}^{x}\mathrm{d}u\int_{-\infty}^{+\infty}p(u,v)\mathrm{d}v=\int_{-\infty}^{x}\left[\int_{-\infty}^{+\infty}p(u,y)\mathrm{d}y\right]\mathrm{d}u$$

从而X的边缘概率密度为

$$p_X(x)=F_X'(x)=\int_{-\infty}^{+\infty}p(x,y)\mathrm{d}y \tag{3-5}$$

同理，Y的边缘概率密度为

$$p_Y(y)=F_Y'(y)=\int_{-\infty}^{+\infty}p(x,y)\mathrm{d}x \tag{3-6}$$

用式（3-5）求$p_X(x)$时，在积分中视x为参数，而用式（3-6）求$p_Y(y)$时，视y为参数。

【例3-6】 设(X,Y)服从二维正态分布$N(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)$，其联合概率密度为

$$p(x,y)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2-\frac{2\rho(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2}+\left(\frac{y-\mu_2}{\sigma_2}\right)^2\right]\right\},-\infty<x,y<+\infty$$

求边缘概率密度 $p_X(x)$， $p_Y(y)$。

解：$p_X(x)=\int_{-\infty}^{+\infty}p(x,y)\mathrm{d}y$

作变量代换，令 $t=\frac{1}{\sqrt{1-\rho^2}}\left(\frac{y-\mu_2}{\sigma_2}-\rho\frac{x-\mu_1}{\sigma_1}\right)$， $\mathrm{d}t=\frac{1}{\sigma_2\sqrt{1-\rho^2}}\mathrm{d}y$，则

$$-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2-\frac{2\rho(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2}+\left(\frac{y-\mu_2}{\sigma_2}\right)^2\right]$$

$$=-\frac{1}{2(1-\rho^2)}\left[\left(\frac{y-\mu_2}{\sigma_2}-\rho\frac{x-\mu_1}{\sigma_1}\right)^2+(1-\rho^2)\left(\frac{x-\mu_1}{\sigma_1}\right)^2\right]$$

$$=-\frac{(x-\mu_1)^2}{2\sigma_1^2}-\frac{t^2}{2}$$

从而

$$p_X(x)=\frac{1}{2\pi\sigma_1}\mathrm{e}^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}\int_{-\infty}^{+\infty}\mathrm{e}^{-\frac{t^2}{2}}\mathrm{d}t$$

利用 $\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{t^2}{2}}\mathrm{d}t=1$，得 $\int_{-\infty}^{+\infty}\mathrm{e}^{-\frac{t^2}{2}}\mathrm{d}t=\sqrt{2\pi}$。于是

$$p_X(x)=\frac{1}{\sqrt{2\pi}\sigma_1}\mathrm{e}^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}，\quad -\infty<x,y<+\infty$$

即 $X\sim N(\mu_1,\sigma_1^2)$

同理，可得

$$p_Y(y)=\frac{1}{\sqrt{2\pi}\sigma_2}\mathrm{e}^{-\frac{(y-\mu_2)^2}{2\sigma_2^2}}，\quad -\infty<x,y<+\infty$$

即 $Y\sim N(\mu_2,\sigma_2^2)$

这个例子说明了二维正态分布的一个重要性质：二维正态分布的边缘分布仍是正态分布。还看到这些边缘分布都与参数 ρ 无关，亦即对于给定的 $\mu_1,\mu_2,\sigma_1^2,\sigma_2^2$，不同的 ρ 对应着不同的二维正态分布，但它们的边缘分布都是相同的。因此，这又一次说明了由边缘分布不能确定 (X,Y) 的联合分布。

要特别注意的是，若 (X,Y) 服从二维正态分布，则 X,Y 必定服从一维正态分布，反之却不一定成立。

综上所述，可以得出结论：联合分布决定边缘分布，但一般仅由边缘分布不能决定联合分布。不过，在一定条件下，由边缘分布也能决定联合分布。

3.2.2 条件分布

本节将首先利用随机事件的条件概率讲解离散型随机变量的条件分布，然后讲解连续型随机变量的条件分布。



1．离散型随机变量的条件分布律

设 (X,Y) 为二维离散型随机变量，联合分布律为

$$P\{X=x_i,Y=y_j\}=p_{ij}，\quad i,j=1,2,\cdots$$

则 X 的边缘分布律为

$$P\{X=x_i\}=\sum_{j=1}^{\infty}p_{ij}=p_{i\cdot}，\quad i=1,2,\cdots$$

Y 的边缘分布律为

$$P\{Y=y_j\}=\sum_{i=1}^{\infty}p_{ij}=p_{\cdot j}，\quad j=1,2,\cdots$$

现在考虑在 $Y=y_j$ 条件下，随机变量 $X=x_i$ 的条件概率，由条件概率公式可得

$$P\{X=x_i\mid Y=y_j\}=\frac{P\{X=x_i,Y=y_j\}}{P\{Y=y_j\}}=\frac{p_{ij}}{p_{\cdot j}}，\quad i=1,2,\cdots$$

易知上述条件概率满足分布律的性质：

1） $P\{X=x_i\mid Y=y_j\}\geqslant 0$。

2） $\sum_{i=1}^{\infty}P\{X=x_i\mid Y=y_j\}=\sum_{i=1}^{\infty}\frac{p_{ij}}{p_{\cdot j}}=\frac{1}{p_{\cdot j}}\sum_{i=1}^{\infty}p_{ij}=\frac{p_{\cdot j}}{p_{\cdot j}}=1$

于是引入下面的定义。

【例 3-7】 对固定的 y_j，若 $P\{Y=y_j\}>0$，则称

$$P\{X=x_i\mid Y=y_j\}=\frac{p_{ij}}{p_{\cdot j}}，\quad i=1,2,\cdots$$

为在 $Y=y_j$ 条件下，随机变量 X 的条件分布律。

同样，对固定的 x_i，若 $P\{X=x_i\}>0$，则称

$$P\{Y=y_j\mid X=x_i\}=\frac{p_{ij}}{p_{i\cdot}}，\quad j=1,2,\cdots$$

为在 $X=x_i$ 条件下，随机变量 Y 的条件分布律。

由此可见，求离散型随机变量的条件分布律无疑是条件概率公式的推广。

2．连续型随机变量的条件概率密度

设 (X,Y) 为二维连续型随机变量，其联合概率密度为 $p(x,y)$。因为对于任意的实数 x,y，都有 $P\{X=x\}=0$，$P\{Y=y\}=0$，因此不能像离散型随机变量那样引入条件分布。下面先用极限的方法来导出条件分布函数。

定义 3-5 给定 y，设对于任意给定的正数 ε，$P\{y-\varepsilon<Y\leqslant y+\varepsilon\}>0$，且对于任意实数 x，极限 $\lim\limits_{\varepsilon\to 0^+}P\{X\leqslant x\mid y-\varepsilon<Y\leqslant y+\varepsilon\}$ 存在，则称此极限为 $Y=y$ 条件下，随机变量 X 的

条件分布函数，记为 $F_{X|Y}(x|y)$ 或 $P\{X \leqslant x | y-\varepsilon < Y \leqslant y+\varepsilon\}$，即

$$F_{X|Y}(x|y) = P\{X \leqslant x | Y=y\} = \lim_{\varepsilon \to 0^+} P\{X \leqslant x | y-\varepsilon < Y \leqslant y+\varepsilon\} \tag{3-7}$$

类似地，可定义在 $X=x$ 条件下，随机变量 Y 的条件分布函数为

下面推导在 $Y=y$ 条件下，随机变量 X 的条件分布函数。

$$\begin{aligned} F_{X|Y}(x|y) &= \lim_{\varepsilon \to 0^+} P\{X \leqslant x | y-\varepsilon < Y \leqslant y+\varepsilon\} \\ &= \lim_{\varepsilon \to 0^+} \frac{P\{X \leqslant x, y-\varepsilon < Y \leqslant y+\varepsilon\}}{P\{y-\varepsilon < Y \leqslant y+\varepsilon\}} = \lim_{\varepsilon \to 0^+} \frac{F(x,y+\varepsilon)-F(x,y-\varepsilon)}{F_Y(y+\varepsilon)-F_Y(y-\varepsilon)} \\ &= \lim_{\varepsilon \to 0^+} \frac{[F(x,y+\varepsilon)-F(x,y-\varepsilon)]/2\varepsilon}{F_Y(y+\varepsilon)-F_Y(y-\varepsilon)/2\varepsilon} = \frac{\dfrac{\partial F(x,y)}{\partial y}}{\dfrac{\mathrm{d}F_Y(y)}{\mathrm{d}y}} \end{aligned} \tag{3-8}$$

因为

$$F(x,y) = \int_{-\infty}^{y} \mathrm{d}v \int_{-\infty}^{x} p(u,v)\mathrm{d}u$$

所以

$$\frac{\partial F(x,y)}{\partial y} = \int_{-\infty}^{x} p(u,y)\mathrm{d}u$$

又因为

$$\frac{\mathrm{d}F_Y(y)}{\mathrm{d}y} = p_Y(y)$$

于是由式（3-8），得

$$F_{X|Y}(x|y) = \frac{\int_{-\infty}^{x} p(u,y)\mathrm{d}u}{p_Y(y)} = \int_{-\infty}^{x} \frac{p(u,y)}{p_Y(y)}\mathrm{d}u \tag{3-9}$$

记 $p_{X|Y}(x|y)$ 表示在 $Y=y$ 条件下，X 的条件概率密度，则

$$p_{X|Y}(x|y) = \frac{f(x,y)}{p_Y(y)} \tag{3-10}$$

同理

$$p_{Y|X}(y|x) = \frac{f(x,y)}{p_X(x)} \tag{3-11}$$

为在 $X=x$ 条件下，Y 的条件概率密度。

由式（3-10）与式（3-11），又得到

$$p(x,y) = p_X(x)p_{Y|X}(y|x) = p_Y(y)p_{X|Y}(x|y) \tag{3-12}$$

从而

$$p_X(x) = \int_{-\infty}^{+\infty} p(x,y)\mathrm{d}y = \int_{-\infty}^{+\infty} p_Y(y)p_{X|Y}(x|y)\mathrm{d}y \tag{3-13}$$

$$p_{Y|X}(y|x) = \frac{p(x,y)}{p_X(x)} = \frac{p_Y(y)p_{X|Y}(x|y)}{\int_{-\infty}^{+\infty} p_Y(y)p_{X|Y}(x|y)\mathrm{d}y} \tag{3-14}$$

【例 3-8】 对于二维连续型随机变量 (X,Y)，已知 Y 的边缘概率密度为



$$p_Y(y)=\begin{cases}6y(1-y), & 0<y<1\\ 0, & 其他\end{cases}$$

且在 $Y=y$（$0<y<1$）条件下，X 的条件概率密度为

$$p_{X|Y}(x\,|\,y)=\begin{cases}\dfrac{1}{1-y}, & y<x<1\\ 0, & 其他\end{cases}$$

求 X 的边缘概率密度 $p_X(x)$ 及 $P\left\{X<\dfrac{1}{2}\right\}$。

解： (X,Y) 的联合概率密度为

$$p(x,y)=p_Y(y)p_{X|Y}(x\,|\,y)=\begin{cases}6y, & 0<y<1, y<x<1\\ 0, & 其他\end{cases}$$

$G=\{(x,y)\,|\,p(x,y)>0\}=\{(x,y)\,|\,0<y<1, y<x<1\}$（见图 3-1 中的阴影部分）。

$$p_X(x)=\int_{-\infty}^{+\infty}p(x,y)\mathrm{d}y=\begin{cases}\displaystyle\int_0^x 6y\mathrm{d}y=3x^2, & 0<x<1\\ 0, & 其他\end{cases}$$

$$P\left\{X<\frac{1}{2}\right\}=\int_{-\infty}^{\frac{1}{2}}p_X(x)\mathrm{d}x=\int_0^{\frac{1}{2}}3x^2\mathrm{d}x=\frac{1}{8}$$

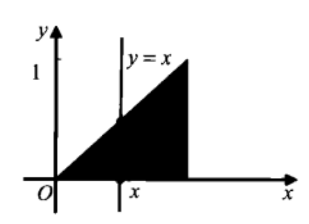


图 3-1　例 3-8 的阴影图

3.2.3 二维正态分布

定义 3-6　如果二维连续型随机向量 (X,Y) 的概率密度为

$$p(x,y)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\mathrm{e}^{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_1)^2}{\sigma_1^2}-2\rho\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2}+\frac{(y-\mu_2)^2}{\sigma_2^2}\right]} \tag{3-15}$$

其中，$-\infty<x<+\infty, -\infty<y<+\infty; \mu_1,\mu_2,\sigma_1,\sigma_2,\rho$ 均为常数，且 $\sigma_1>0, \sigma_2>0, -1<\rho<1$，则称 (X,Y) 服从参数 $\mu_1,\mu_2,\sigma_1,\sigma_2,\rho$ 的二维正态分布。

【例 3-9】　求服从参数 $\mu_1,\mu_2,\sigma_1,\sigma_2,\rho$ 的二维正态分布的边缘概率密度。

解： 由式（3-5）知

$$p_X(x)=\int_{-\infty}^{+\infty}p(x,y)\mathrm{d}y$$

由于

$$\frac{(y-\mu_2)^2}{\sigma_2^2}-2\rho\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2}=\left(\frac{y-\mu_2}{\sigma_2}-\rho\frac{x-\mu_1}{\sigma_1}\right)^2-\rho^2\frac{(x-\mu_1)^2}{\sigma_1^2}$$

于是

$$p_X(x)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\mathrm{e}^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}\int_{-\infty}^{+\infty}\mathrm{e}^{-\frac{1}{2(1-\rho^2)}\left(\frac{y-\mu_2}{\sigma_2}-\rho\frac{x-\mu_1}{\sigma_1}\right)^2}\mathrm{d}y$$

令

$$t=\frac{1}{\sqrt{1-\rho^2}}\left(\frac{y-\mu_2}{\sigma_2}-\rho\frac{x-\mu_1}{\sigma_1}\right)$$

则有 $\mathrm{d}t=\frac{1}{\sigma_2\sqrt{1-\rho^2}}\mathrm{d}y$，且当 $y\to+\infty$ 时，$t\to+\infty$；当 $y\to-\infty$ 时，$t\to-\infty$。因此

$$p_X(x)=\frac{1}{2\pi\sigma_1}\mathrm{e}^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}\int_{-\infty}^{+\infty}\mathrm{e}^{-\frac{t^2}{2}}\mathrm{d}t=\frac{1}{\sigma_1\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2\sigma_1^2}(x-\mu_1)^2}\quad(-\infty<x<+\infty)$$

同理，可得

$$p_Y(x)=\frac{1}{\sigma_2\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2\sigma_2^2}(y-\mu_2)^2}\quad(-\infty<y<+\infty)$$

例 3-9 的结果表明，二维正态分布的两个边缘分布都是一维正态分布；而且二维正态分布 5 个参数中的 μ_1,μ_2 和 σ_1^2,σ_2^2 分别是其两个边缘概率密度的均值和方差。除此之外，还可说明式（3-15）所给出的 $p(x,y)$ 是一个二维连续型随机向量的联合概率密度。这是因为

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}p(x,y)\mathrm{d}x\mathrm{d}y=\int_{-\infty}^{+\infty}\left[\int_{-\infty}^{+\infty}p(x,y)\mathrm{d}y\right]\mathrm{d}x=\int_{-\infty}^{+\infty}\frac{1}{\sigma_1\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2\sigma_1^2}(x-\mu_1)^2}\mathrm{d}x=1$$

作为本节的结束，再给出一个关于二维正态分布的重要结论，其证明从略，在应用中可以直接运用。

如果随机向量 (X,Y) 服从参数 $\mu_1,\mu_2,\sigma_1,\sigma_2,\rho$ 的二维正态分布，则 X 与 Y 相互独立的充分必要条件为

$$\rho=0$$

3.3 随机向量函数的分布

在前面章节中，介绍了一个随机变量函数的分布问题，即已知随机变量 X 的分布，研究 X 的函数 $Y=f(X)$ 的分布问题。在本节中，将讲解二维随机向量函数的分布问题。具体地讲，就是已知二维随机向量 (X,Y) 的联合分布，求其函数 $Z=f(X,Y)$。

3.3.1 二维随机向量函数的概念

定义 3-7 设 $f(x,y)$ 是定义在二维随机向量 (X,Y) 一切可能取值集合上的二元函数。如果二维随机向量 (X,Y) 取值 (x,y) 时，随机变量 Z 取值 $z=f(x,y)$，则称 Z 为二维随机向量 (X,Y) 的函数，记为 $Z=f(X,Y)$。

设二维连续型随机向量(X,Y)的联合密度为$p(x,y)$，$Z=f(X,Y)$的分布函数记为$F_Z(z)$，则有

$$F_Z(z)=P(Z\leqslant z)=P[f(X,Y)\leqslant z]=\iint\limits_{f(x,y)\leqslant z}p(x,y)\mathrm{d}x\mathrm{d}y \qquad (3-16)$$



这是二维连续型随机向量函数的分布计算公式。下面就几个具体的二维随机向量函数进行讲解。

3.3.2 函数分布

1. $Z=X+Y$的分布

设随机向量(X,Y)的联合密度为$p(x,y)$，则由式（3-16）知

$$F_Z(z)=P(X+Y\leqslant z)=\iint\limits_{x+y\leqslant z}p(x,y)\mathrm{d}x\mathrm{d}y$$

这里积分区域$D=\{(x,y)\,|\,x+y\leqslant z\}$是$xOy$平面上的一个区域，如图3-2所示。

利用二重积分与累积分的关系，并令$y=u-x$，得

$$\iint\limits_{x+y\leqslant z}p(x,y)\mathrm{d}x\mathrm{d}y=\int_{-\infty}^{+\infty}\mathrm{d}x\int_{-\infty}^{z-x}p(x,y)\mathrm{d}y$$

$$=\int_{-\infty}^{+\infty}\mathrm{d}x\int_{-\infty}^{-z}p(x,u-x)\mathrm{d}u=\int_{-\infty}^{-z}\mathrm{d}u\int_{-\infty}^{+\infty}p(x,u-x)\mathrm{d}x$$

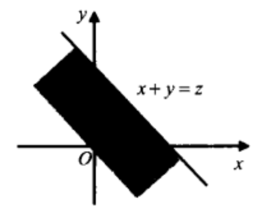


图3-2　积分区域

因此

$$F_Z(z)=\int_{-\infty}^{-z}\left[\int_{-\infty}^{+\infty}p(x,u-x)\mathrm{d}x\right]\mathrm{d}u$$

再将此式两边对z求导，即得

$$F_Z'(z)=p_Z(z)=\int_{-\infty}^{+\infty}p(x,z-x)\mathrm{d}x \qquad (3-17)$$

由X,Y的对称性知，$p_Z(z)$还可表示为

$$p_Z(z)=\int_{-\infty}^{+\infty}p(z-y,y)\mathrm{d}y \qquad (3-18)$$

其推导从略。

【例3-10】 设X与Y相互独立，且$X\sim N(\mu_1,\sigma_1^2)$，$Y\sim N(\mu_2,\sigma_2^2)$，求$Z=X+Y$的密度$p_Z(z)$。

解： 由题设知(X,Y)的联合密度为

$$p(x,y)=\frac{1}{2\pi\sigma_1\sigma_2}\mathrm{e}^{-\frac{1}{2}\left[\frac{(x-\mu_1)^2}{\sigma_1^2}+\frac{(y-\mu_2)^2}{\sigma_2^2}\right]},\quad -\infty<x<+\infty,\quad -\infty<y<+\infty$$

于是根据式（3-17），即得

$$p_Z(z)=\int_{-\infty}^{+\infty}p(x,z-x)\mathrm{d}x=\int_{-\infty}^{+\infty}\frac{1}{2\pi\sigma_1\sigma_2}\mathrm{e}^{-\frac{1}{2}\left[\frac{(x-\mu_1)^2}{\sigma_1^2}+\frac{(y-\mu_2)^2}{\sigma_2^2}\right]}\mathrm{d}x$$

又由于

$$\frac{(x-\mu_1)^2}{\sigma_1^2}+\frac{(z-x-\mu_2)^2}{\sigma_2^2}$$

$$=\frac{1}{\sigma_1\sigma_2}\{\sigma_2^2(x-\mu_1)^2+\sigma_1^2[(z-\mu_1-\mu_2)-(x-\mu_1)]^2\}$$

$$=\frac{1}{\sigma_1\sigma_2}[(\sigma_2^2+\sigma_1^2)(x-\mu_1)^2+\sigma_1^2[(z-\mu_1-\mu_2)^2-2\sigma_1^2(x-\mu_1)(z-\mu_1-\mu_2)]$$

$$=\frac{1}{\sigma_1\sigma_2}\left[\sqrt{\sigma_2^2+\sigma_1^2}(x-\mu_1)-\frac{\sigma_1^2}{\sqrt{\sigma_2^2+\sigma_1^2}}(z-\mu_1-\mu_2)\right]^2+\frac{1}{\sigma_1+\sigma_2}(z-\mu_1-\mu_2)^2$$

所以令 $t=\sqrt{\sigma_2^2+\sigma_1^2}(x-\mu_1)$，则有 $\mathrm{d}x=\frac{1}{\sqrt{\sigma_2^2+\sigma_1^2}}\mathrm{d}t$，且当 $x\to+\infty$ 时，$t\to+\infty$；当 $x\to-\infty$，$t\to-\infty$。因此

$$p_Z(z)=\int_{-\infty}^{+\infty}\frac{1}{2\pi\sigma_1\sigma_2}\mathrm{e}^{-\frac{1}{2}\left[\frac{(x-\mu_1)^2}{\sigma_1^2}+\frac{(y-\mu_2)^2}{\sigma_2^2}\right]}\mathrm{d}x=\frac{1}{\sqrt{\sigma_2^2+\sigma_1^2}\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2(\sigma_2^2+\sigma_1^2)}[z-(\mu_1+\mu_2)^2]}$$

$$\int_{-\infty}^{+\infty}\frac{1}{\sigma_1\sigma_2\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2(\sigma_2^2+\sigma_1^2)^2}\left[t-\frac{\sigma_1^2}{\sqrt{\sigma_2^2+\sigma_1^2}}z-(\mu_1+\mu_2)\right]^2}\mathrm{d}t=\frac{1}{\sqrt{\sigma_2^2+\sigma_1^2}\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2(\sigma_2^2+\sigma_1^2)}[z-(\mu_1+\mu_2)^2]}$$

这表明 $Z=X+Y\sim N(\mu_1+\mu_2,\sigma_2^2+\sigma_1^2)$，即 $Z=X+Y$ 也服从正态分布，其均值与方差都是 X 与 Y 的均值之和与方差之和。

2. $Z=\sqrt{X^2+Y^2}$ 的分布

设 X 与 Y 相互独立，且 X 和 Y 的分布函数及密度分别为 $F_X(x)$，$F_Y(y)$ 及 $p_X(x)$，$p_Y(y)$。求 $Z=\max(X,Y)$ 的分布函数 $F_{\max}(z)$ 及密度 $p_{\max}(z)$。

因为 X 与 Y 相互独立，所以

$$P(X\leqslant z,Y\leqslant z)=p(X\leqslant z)p(Y\leqslant z)=F_X(z)F_Y(z)$$

又由于 $Z=\max(X,Y)$ 不大于 z 等价于 X 和 Y 都不大于 z，即

$$P(Z\leqslant z)=P(X\leqslant z,Y\leqslant z)$$

因此，$Z=\max(X,Y)$ 的分布函数为

$$F_{\max}=P(Z\leqslant z)=F_X(z)F_Y(z)\tag{3-19}$$

进而可知 $Z=\max(X,Y)$ 的密度为

$$p_{\max}(z)=F'_{\max}(z)=F'_X(z)F_Y(z)+F_X(z)F'_Y(z)=p_X(z)F_Y(z)+F_X(z)p_Y(z)\tag{3-20}$$

【例 3-11】 设系统 L 由两个相互独立的子系统 L_1，L_2 联接而成。联接的方式分别为 1）串联（见图 3-3a）。2）并联（见图 3-3b）。3）备用（备用系统是指当系统 L_1 损坏时，系统 L_2 开始工作，见图 3-3c）。

现已知系统 L_1、L_2 的寿命分别为 X 和 Y，其密度分别为

$$p_X(x)=\begin{cases}\alpha\mathrm{e}^{-\alpha x}, & x>0\\ 0, & x\leqslant 0\end{cases},\quad p_Y(y)=\begin{cases}\beta\mathrm{e}^{-\beta y}, & y>0\\ 0, & y\leqslant 0\end{cases}$$

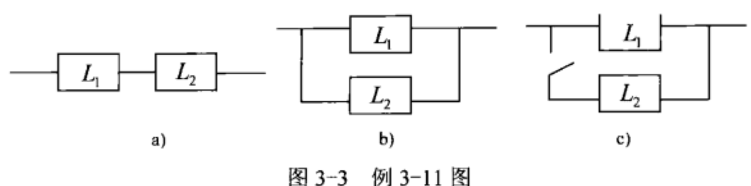




图 3-3　例 3-11 图

a) 串联　b) 并联　c) 备用系统

其中，$\alpha>0,\beta>0$ 且 $\alpha\neq\beta$。试分别就以上 3 种联接方式求出系统 L 的寿命 Z 的密度。

解：根据题设由计算知 X，Y 的分布函数分别为

$$F_X(x)=\begin{cases}1-\alpha e^{-\alpha y}, & x>0\\ 0, & x\leqslant 0\end{cases},\quad F_Y(y)=\begin{cases}1-\beta e^{-\beta y}, & y>0\\ 0, & y\leqslant 0\end{cases}$$

1）串联：由于当 L_1，L_2 中有一个损坏时，系统 L 就停止工作，所以这时 L 的寿命 $Z=\min(X,Y)$。

如果记 $Z=\min(X,Y)$ 的分布函数为 $F_{\min}(z)$，则有

$$\begin{aligned}F_{\min}(z)&=P(Z\leqslant z)=1-P(Z>z)=1-P(X>z,Y>z)=1-P(X>z)P(Y>z)\\&=1-[1-F_X(x)][1-F_Y(y)]\end{aligned}$$

即 $Z=\min(X,Y)$ 的概率密度为

$$F_{\min}(z)=\begin{cases}1-e^{-(\alpha+\beta)z}, & z>0\\ 0, & z\leqslant 0\end{cases}$$

于是 $Z=\min(X,Y)$ 的概率密度为

$$p_{\min}(z)=F'_{\min}(z)=\begin{cases}(\alpha+\beta)e^{-(\alpha+\beta)z}, & z>0\\ 0, & z\leqslant 0\end{cases}$$

2）并联：由于当且仅当 L_1，L_2 都损坏时，系统 L 才停止工作，所以这时 L 的寿命为 $Z=\max(X,Y)$。

于是利用式（3-19）得 $Z=\max(X,Y)$ 的分布函数为

$$F_{\max}(z)=F_X(x)F_Y(y)=\begin{cases}(1-e^{-\alpha z})(1-e^{-\beta z}), & z>0\\ 0, & z\leqslant 0\end{cases}$$

再利用式（3-20）得 $Z=\max(X,Y)$ 的概率密度为

$$p_{\max}(z)=\begin{cases}\alpha e^{-\alpha z}+\beta e^{-\beta z}-(\alpha+\beta)e^{-(\alpha+\beta)z}, & z>0\\ 0, & z\leqslant 0\end{cases}$$

3）备用：由于这时当系统 L_1 损坏时，系统 L_2 才开始工作，此整个系统 L 的寿命 Z 是 L_1 与 L_2 两者寿命之和，即 $Z=X+Y$。

于是利用式（3-18）得 $Z=X+Y$ 的概率密度为

$$p(z)=\int_{-\infty}^{+\infty}p_X(z-y)p_Y(y)\mathrm{d}y$$

由于 $p_Y(y)$ 仅当 $y>0$ 时是非零值，而 $p_X(z-y)$ 仅当 $z-y>0$（即 $y<z$）时是非零值，所以上式右端的被积函数 $p_X(z-y)p_Y(y)$ 仅当 $0<y<z$ 时是非零值，因此

$$p_z(z)=\int_0^z \alpha e^{-\alpha z}\beta e^{-\beta z}dy=\alpha\beta e^{-\alpha z}\int_0^z e^{-(\beta-\alpha)y}dy=\frac{\alpha\beta}{\beta-\alpha}(e^{-\alpha z}-e^{-\beta z})$$

又由于当 $z\leqslant 0$ 时，$p(z)=0$，于是 $Z=X+Y$ 的概率密度为

$$p(z)=\begin{cases}\dfrac{\alpha\beta}{\beta-\alpha}(e^{-\alpha z}-e^{-\beta z}), & z>0\\ 0, & z\leqslant 0\end{cases}$$

3. $Z=\sqrt{X^2+Y^2}$ 的分布

由上述内容可知，为求随机向量函数 $Z=f(X,Y)$ 的概率密度，先要求其分布函数

$$F_Z=P\{f(X,Y)<z\}$$

而在求其分布函数的过程中，需要用到式（3-16）。下面就利用式（3-16），再计算一个重要的例子。

【例 3-12】 设 X 与 Y 相互独立，服从相同的分布 $N(0,\sigma^2)$。求 $Z=\sqrt{X^2+Y^2}$ 的概率密度。

解：由题设知，X 与 Y 的概率密度分别为

$$p_X(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{x^2}{2\sigma^2}},\quad -\infty<x<+\infty$$

$$p_Y(y)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{y^2}{2\sigma^2}},\quad -\infty<y<+\infty$$

所以随机向量 (X,Y) 的联合概率密度为

$$p(x,y)=\frac{1}{2\pi\sigma^2}e^{-\frac{1}{2\sigma^2}(x^2+y^2)},\quad -\infty<x<+\infty,-\infty<y<+\infty$$

由于 $Z=\sqrt{X^2+Y^2}$ 只取非负值，所以当 $z<0$ 时，其分布函数 $F_Z(z)=0$；而当 $z\geqslant 0$ 时，利用式（3-16），即得

$$F_Z(z)=P(\sqrt{X^2+Y^2}\leqslant z)=\iint\limits_{\sqrt{X^2+Y^2}\leqslant z}\frac{1}{2\pi\sigma^2}e^{-\frac{1}{2\sigma^2}(x^2+y^2)}dxdy$$

引入极坐标计算上述二重积分，即得

$$F_Z(z)=\frac{1}{2\pi\sigma^2}\int_0^{2\pi}d\theta\int_0^z e^{-\frac{r^2}{2\sigma^2}}rdr=\frac{1}{2\pi\sigma^2}2\pi\sigma^2(1-e^{-\frac{z^2}{2\sigma^2}})=1-e^{-\frac{z^2}{2\sigma^2}}$$

于是 $Z=\sqrt{X^2+Y^2}$ 的分布函数为

$$F_Z(z)=\begin{cases}1-e^{-\frac{z^2}{2\sigma^2}}, & z\geqslant 0\\ 0, & z<0\end{cases}$$

由此可得 $Z=\sqrt{X^2+Y^2}$ 的概率密度为

$$p_Z(z)=\begin{cases}\dfrac{z}{\sigma^2}e^{-\frac{z^2}{2\sigma^2}}, & z\geqslant 0\\ 0, & z<0\end{cases}$$

这就是参数为$\sigma(\sigma>0)$的 Rayleigh 分布。

3.4 二维随机向量的数字特征



本节将介绍二维随机向量的数字特征。

3.4.1 数学期望

定理 3-1 设(X,Y)为二维随机变量，$g(x,y)$为二元连续函数。

1）若(X,Y)为二维离散型随机变量，其联合分布律为

$$P\{X=x_i,Y=y_j\}=p_{ij}，\quad i,j=1,2,\cdots$$

且级数$\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}g(x_i,y_j)p_{ij}$绝对收敛，则随机变量函数$g(X,Y)$的数学期望为

$$E[g(X,Y)]=\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}g(x_i,y_j)p_{ij} \tag{3-21}$$

2）若(X,Y)为二维连续型随机变量，其联合概率密度为$p(x,y)$，且广义积分$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}g(x,y)p(x,y)\mathrm{d}x\mathrm{d}y$绝对收敛，则随机变量函数$g(X,Y)$的数学期望为

$$E[g(X,Y)]=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}g(x,y)p(x,y)\mathrm{d}x\mathrm{d}y \tag{3-22}$$

注意到，若(X,Y)为二维连续型随机变量，其联合概率密度为$p(x,y)$。由式（3-22）得

$$E[g(X)]=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}g(x)p(x,y)\mathrm{d}x\mathrm{d}y=\int_{-\infty}^{+\infty}g(x)\left[\int_{-\infty}^{+\infty}p(x,y)\mathrm{d}y\right]\mathrm{d}x=\int_{-\infty}^{+\infty}g(x)p_X(x)\mathrm{d}x$$

对于二维连续型随机变量，计算$E[g(X)]$可用

$$E[g(X)]=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}g(x)p(x,y)\mathrm{d}x\mathrm{d}y \tag{3-23}$$

或

$$E[g(X)]=\int_{-\infty}^{+\infty}g(x)p_X(x)\mathrm{d}x \tag{3-24}$$

如果用后者计算，需先由联合概率密度$p(x,y)$计算X的边缘概率密度$p_X(x)$，不如用式（3-23）方便。当(X,Y)为二维离散型随机变量时，由于求边缘分布律不复杂，所以一般用式（3-24）求。

【例 3-13】 设(X,Y)的联合概率密度为

$$p(x,y)=\begin{cases}3x, & 0<x<1,0<y<x\\ 0, & 其他\end{cases}$$

求：1）$E(X)$，$E(X^2)$。2）$E(Y)$。3）$E(X+Y)$。4）$E(XY)$。

解：用定理 3-1 计算。

1）$E(X)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}xp(x,y)\mathrm{d}x\mathrm{d}y=\int_0^1\mathrm{d}x\int_0^x 3x^2\mathrm{d}y=\int_0^1 3x^3\mathrm{d}x=\frac{3}{4}$

$$E(X^2)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}x^2p(x,y)\mathrm{d}x\mathrm{d}y=\int_0^1\mathrm{d}x\int_0^x 3x^3\mathrm{d}y=\int_0^1 3x^4\mathrm{d}x=\frac{3}{5}$$

2） $E(Y)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}yp(x,y)\mathrm{d}x\mathrm{d}y=\int_0^1\mathrm{d}x\int_0^x 3xy\mathrm{d}y=\int_0^1\frac{3}{2}x^3\mathrm{d}x=\frac{3}{8}$

3） $E(X+Y)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(x+y)p(x,y)\mathrm{d}x\mathrm{d}y=\int_0^1\mathrm{d}x\int_0^x 3x(x+y)\mathrm{d}y=\int_0^1\frac{9}{2}x^3\mathrm{d}x=\frac{9}{8}$

4） $E(XY)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}xyp(x,y)\mathrm{d}x\mathrm{d}y=\int_0^1\mathrm{d}x\int_0^x 3x^2y\mathrm{d}y=\int_0^1\frac{3}{2}x^4\mathrm{d}x=\frac{3}{10}$

3.4.2 边缘分布的期望与方差

设二维随机向量(X,Y)的联合概率密度为$p(x,y)$，其关于X,Y的边缘概率密度分别为$p_X(x)$，$p_Y(y)$，由例 3-13 和(X,Y)的联合概率密度$p(x,y)$计算X,Y的期望与方差，得下列公式。

$$E(X)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}xp(x,y)\mathrm{d}x\mathrm{d}y \tag{3-25}$$

$$E(Y)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}yp(x,y)\mathrm{d}x\mathrm{d}y \tag{3-26}$$

$$D(X)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}[x-E(X)]^2p(x,y)\mathrm{d}x\mathrm{d}y \tag{3-27}$$

$$D(Y)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}[y-E(Y)]^2p(x,y)\mathrm{d}x\mathrm{d}y \tag{3-28}$$

【例 3-14】 设二维随机向量(X,Y)的联合概率密度为

$$p(x,y)=\begin{cases}\frac{1}{8}(x+y), & 0\leqslant x\leqslant 2,0\leqslant y\leqslant 2\\ 0, & \text{其他}\end{cases}$$

求：1）$D(X)$。2）$D(Y)$。3）$E(X)$。4）$E(Y)$。

解：利用式（3-25）～式（3-28），即得

1） $E(X)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}xp(x,y)\mathrm{d}x\mathrm{d}y=\int_0^2 x\left[\int_0^2\frac{1}{8}(x+y)\mathrm{d}y\right]\mathrm{d}x$

$$=\int_0^2 x\left[\frac{1}{4}(x+1)\right]\mathrm{d}x=\left.\frac{x^3}{12}+\frac{x^2}{8}\right|_0^2=\frac{7}{6}$$

2） $E(Y)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}yp(x,y)\mathrm{d}x\mathrm{d}y=\int_0^2 y\left[\int_0^2\frac{1}{8}(x+y)\mathrm{d}x\right]\mathrm{d}y=\int_0^2 y\left[\frac{1}{4}(y+1)\right]\mathrm{d}y=\frac{7}{6}$

3） $D(X)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\left(x-\frac{7}{6}\right)^2p(x,y)\mathrm{d}x\mathrm{d}y=\int_0^2\left(x-\frac{7}{6}\right)^2\left[\int_0^2\frac{1}{8}(x+y)\mathrm{d}y\right]\mathrm{d}x$

$$=\int_0^2\left(x-\frac{7}{6}\right)^2\left[\frac{1}{4}(x+1)\right]\mathrm{d}x=\frac{11}{36}$$

4） $D(Y)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\left(y-\frac{7}{6}\right)^2p(x,y)\mathrm{d}x\mathrm{d}y=\int_0^2\left(y-\frac{7}{6}\right)^2\left[\int_0^2\frac{1}{8}(x+y)\mathrm{d}x\right]\mathrm{d}y$

$$= \int_0^2 \left(y-\frac{7}{6}\right)^2 \left[\frac{1}{4}(y+1)\right] dy = \frac{11}{36}$$

3.4.3 协方差



定义 3-8 设(X,Y)为二维随机变量，若$E\{[X-E(X)][Y-E(Y)]\}$存在，则称它是随机变量X与Y的协方差，记为$\text{Cov}(X,Y)$，即

$$\text{Cov}(X,Y)=E\{[X-E(X)][Y-E(Y)]\} \tag{3-29}$$

显然，有$\text{Cov}(X,Y)=D(X)$。

由式（3-29）计算协方差$\text{Cov}(X,Y)$，实际上就是计算二维随机变量(X,Y)的函数$g(X,Y)=E\{[X-E(X)][Y-E(Y)]\}$的数学期望。

若(X,Y)为二维离散型随机变量，联合分布律为

$$P\{X=x_i,Y=y_j\}=p_{ij}\text{，}\quad i,j=1,2,\cdots$$

则由式（3-21），得

$$\text{Cov}(X,Y)=\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}[x_i-E(X)][y_j-E(Y)]p_{ij} \tag{3-30}$$

若(X,Y)为二维连续型随机变量，联合概率密度为$p(x,y)$，则由式（3-22），得

$$\text{Cov}(X,Y)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}[x-E(X)][y-E(Y)]\text{d}x\text{d}y \tag{3-31}$$

为便于计算协方差$\text{Cov}(X,Y)$，常采用公式

$$\text{Cov}(X,Y)=E(X,Y)-E(X)E(Y) \tag{3-32}$$

证明：
$$\begin{aligned}\text{Cov}(X,Y)&=E\{[X-E(X)][Y-E(Y)]\}\\&=E[XY-XE(Y)-YE(X)+E(X)E(Y)]\\&=E(XY)-E(X)E(Y)-E(X)E(Y)+E(X)E(Y)\\&=E(XY)-E(X)E(Y)\end{aligned}$$

在例 3-13 中，对于二维连续型随机变量(X,Y)，已求出$E(X)=\frac{3}{4}$，$E(Y)=\frac{3}{8}$，$E(XY)=\frac{3}{10}$，从而X与Y的协方差

$$\text{Cov}(X,Y)=E(XY)-E(X)E(Y)=\frac{3}{10}-\frac{3}{4}\times\frac{3}{8}=\frac{3}{160}$$

协方差具有下列性质：

1）$\text{Cov}(X,Y)=\text{Cov}(Y,X)$。

2）$\text{Cov}(aX+c,bY+d)=ab\text{Cov}(X,Y)$，这里$a,b,c,d$均为常数。

推论 3-1 $\text{Cov}(aX,bY)=ab\text{Cov}(X,Y)$，$a,b$为常数。

3）$\text{Cov}(X_1,X_2,Y)=\text{Cov}(X_1,Y)+\text{Cov}(X_2,Y)$。

推论 3-2 $\text{Cov}(X_1+X_2,Y_1+Y_2)=\text{Cov}(X_1,Y_1)+\text{Cov}(X_1,Y_2)+\text{Cov}(X_2,Y_1)+\text{Cov}(X_2,Y_2)$。

4）若X与Y相互独立，则$\text{Cov}(X,Y)=0$。

5）$D(X+Y)=D(X)+D(Y)+2\text{Cov}(X,Y)$。

推论 3-3 $D(X-Y)=D(X)+D(Y)-2\mathrm{Cov}(X,Y)$。

推论 3-4 若 X 与 Y 相互独立，则 $D(X\pm Y)=D(X)+D(Y)$。

MATLAB 中提供了专门求解多元随机变量协方差均值的 cov 函数。参看以下示例。

【例 3-15】 试用 MATLAB 语言产生 4 个满足标准正态分布的随机变量，并求出其协方差矩阵。

分析：用 MATLAB 给出的 randn 函数可以生成一个标准正态分布随机数的矩阵。该矩阵有 4 列，表示 4 个不同的随机数变量。该矩阵有 30000 行，表示每个随机数变量均取 30000 个样本点。这样，由下面的语句可以立即得出这 4 个随机数变量的协方差矩阵。可见，该矩阵是对称矩阵，趋近于理论上的单位矩阵。

其实现的 MATLAB 程序代码如下：

```
>> p=randn(30000,4);
cov(p)
```

运行程序，输出如下：

```
ans =
    1.0064    0.0013    0.0047   -0.0005
    0.0013    1.0040   -0.0009    0.0048
    0.0047   -0.0009    1.0110   -0.0119
   -0.0005    0.0048   -0.0119    0.9948
```

3.4.4 相关系数

1. 相关系数的基本概念

定义 3-9 $D(X)>0$，$D(Y)>0$，则称数值 $\frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Cov}(X)}\sqrt{\mathrm{Cov}(Y)}}$ 为随机变量 X 与 Y 的线性相关系数，简称相关系数，记为 ρ_{XY}，即

$$\rho_{XY}=\frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Cov}(X)}\sqrt{\mathrm{Cov}(Y)}} \tag{3-33}$$

【例 3-16】 设随机向量 (X,Y) 的联合概率密度为

$$p(x,y)=\begin{cases}A, & 0\leqslant x\leqslant 1, 0\leqslant y\leqslant x\\ 0, & \text{其他}\end{cases}$$

求：1）常数 A。2）相关系数 ρ。

解：1）由于

$$1=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}p(x,y)\mathrm{d}x\mathrm{d}y=\int_0^1\mathrm{d}x\int_0^x A\mathrm{d}y=\frac{A}{2}$$

所以 $A=2$。

2）由 1）的结果可知 (X,Y) 的联合概率密度为

$$p(x,y)=\begin{cases}2, & 0\leqslant x\leqslant 1, 0\leqslant y\leqslant x\\ 0, & \text{其他}\end{cases}$$



从而利用式（3-25）～式（3-28）可得

$$E(X)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}xp(x,y)\mathrm{d}x\mathrm{d}y=\int_0^1 x\left(\int_0^x 2\mathrm{d}y\right)\mathrm{d}x=\frac{2}{3}x^3\Big|_0^1=\frac{2}{3}$$

$$E(Y)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}yp(x,y)\mathrm{d}x\mathrm{d}y=\int_0^1 2\left(\int_0^x y\mathrm{d}y\right)\mathrm{d}x=\frac{1}{3}x^3\Big|_0^1=\frac{1}{3}$$

$$\mathrm{Cov}(X)=D(X)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}[x-E(x)^2]p(x,y)\mathrm{d}x\mathrm{d}y=\int_0^1\int_0^x\left(x-\frac{2}{3}\right)^2 p(x,y)\mathrm{d}x\mathrm{d}y$$

$$=\int_0^1 2\left(x-\frac{2}{3}\right)^2\left(\int_0^x\mathrm{d}y\right)\mathrm{d}x=\int_0^1 2\left(x^3-\frac{4}{3}x^2+\frac{4}{9}x\right)\mathrm{d}x=\frac{1}{18}$$

$$\mathrm{Cov}(Y)=D(Y)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}[y-E(Y)]^2p(x,y)\mathrm{d}x\mathrm{d}y=\int_0^1\int_0^x\left(y-\frac{1}{3}\right)^2 p(x,y)\mathrm{d}x\mathrm{d}y=\int_0^1 2\left[\int_0^x\left(y-\frac{1}{3}\right)^2\mathrm{d}y\right]\mathrm{d}x$$

$$=\int_0^1\frac{2}{3}\left(x^3-x^2+\frac{1}{3}x\right)\mathrm{d}x=\frac{1}{18}$$

利用式（3-31）可得

$$\mathrm{Cov}(X,Y)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}[x-E(X)][y-E(Y)]\mathrm{d}x\mathrm{d}y=\int_0^1\int_0^x\left(x-\frac{2}{3}\right)\left(y-\frac{1}{3}\right)p(x,y)\mathrm{d}x\mathrm{d}y$$

$$=\int_0^1 2\int_0^x\left(x-\frac{2}{3}\right)\left[\int_0^x\left(y-\frac{1}{3}\right)\mathrm{d}y\right]\mathrm{d}x=\int_0^1 x\left(x-\frac{2}{3}\right)^2\mathrm{d}x=\int_0^1\left(x^3-\frac{4}{3}x^2+\frac{4}{9}x\right)\mathrm{d}x=\frac{1}{36}$$

从而再利用式（3-33），即得(X,Y)的相关系数为

$$\rho_{XY}=\frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Cov}(X)}\sqrt{\mathrm{Cov}(Y)}}=\frac{\frac{1}{36}}{\sqrt{\frac{1}{18}}\times\sqrt{\frac{1}{18}}}=\frac{1}{2}$$

2．数字特征的简单性质

下面给出数字特征的一些简单性质，其证明从略。

1）$E(X\pm Y)=E(X)\pm E(Y)$。

2）如果X与Y相互独立，则有$E(XY)=E(X)E(Y)$。

3）$D(X\pm Y)=D(X)+D(Y)\pm 2\sigma_{XY}$。

4）如果X与Y相互独立，则有$D(X\pm Y)=D(X)+D(Y)$。

5）如果X与Y相互独立，则有$\sigma_{XY}=0$（或$\rho_{XY}=0$）。

6）$|\rho_{XY}|\leqslant 1$。

随机变量的期望体现了随机变量取值的平均；随机变量的方差刻画了随机变量的取值与其均值的偏离程度；随机向量的协方差则反映了其两个分量之间的联系；随机向量的相关系数也称为标准协方差，它刻画了随机向量两个分量线性关系的近似程度。一般地讲，$|\rho|$越接近于1，两个分量间越近似地有线性关系。

3.4.5 矩与协方差矩阵

在随机变量的数字特征中，除了数学期望、方差、协方差和相关系数外，还有其他的数

字特征。

1．矩

定义 3-10 对于随机变量 X，若 $E(X^k)$ 为 X 的 k 阶原点矩（简称 k 阶矩），若 $E\{[X-E(X)]^k\}$ 为 X 的 k 阶中心矩，则易知 X 的一阶原点矩为 X 的数学期望。X 的一阶中心矩为零，X 的二阶中心矩为 X 的方差。

注意到

$$|X|^{k-1} \leqslant 1+|X|^k \text{，} k=1,2,\cdots$$

事实上，当 $|X| \geqslant 1$ 时，不等式显然成立；当 $|X|<1$ 时，$|X|^{k-1} \leqslant 1$，故 $|X|^{k-1} \leqslant 1+|X|^k$ 成立。

由上式可推出

$$E(|X|^{k-1}) \leqslant E(1+|X|^k) = 1+E(|X|^k)$$

因此，如果高阶矩 $E(|X|^k)<\infty$，则低阶矩 $E(|X|^{k-1})<\infty$，即若 X 的 k 阶矩存在，则 X 的 $k-1$ 阶矩也存在，从而低于 k 的各阶矩都存在。

定义 3-11 对于随机变量 X,Y，若 $E(X^kY^l)$ 存在，则称 $E(X^kY^l)$ 为 X 与 Y 的 $k+l$ 阶混合矩；若 $E\{[X-E(X)]^k[Y-E(Y)]^l\}$ 存在，则称它为 X 与 Y 的 $k+l$ 阶混合中心矩，$k,l=1,2,\cdots$

由定义 3-11 知，协方差 $\mathrm{Cov}(X,Y)$ 为 X 与 Y 的 1+1 阶混合中心矩。

【例 3-17】 求取 Γ 分布 $(\alpha>0,\lambda>0)$ 的原点矩和中心矩，并由前几项结果总结一般规律。

```
%先用下面的 MATLAB 语句求原点矩
>> syms x;
syms a lam positive;
p=lam^a*x^(a-1)/gamma(a)*exp(-lam*x);
for n=1:5
    m=int(x^n*p,x,0,inf);
end
%结果由下面的语句直接给出：
>> syms n;
m=simple(int((x)^n*p,x,0,inf)
m = lam^(-n)*gamma(n+a)/gamma(a)
%通过下面的语句求出中心矩
>> for n=1:7,
    s=simple(int((x-1/lam*a)^n*p,x,0,inf));
end
s
s =6*a*(120+154*a+35*a^2)/lam^7
```

MATLAB 的统计工具箱提供了 moment 函数，可以求出向量 x 的中心高阶矩，但没有直接函数可以求出原点矩。其实，可以用下面的语句求出给定随机向量 x 的 r 阶原点矩与中心矩：

$A_r = \mathrm{sum}(x.\hat{}\,r)/\mathrm{length}(x)$，$B_r = \mathrm{moment}(x,r)$



【例 3-18】 仍考虑例 3-17 中的随机数，可以用下面的语句得出随机数的各阶矩。

其实现的 MATLAB 程序代码如下：

```
>> A=[];B=[];
p=normrnd(0.5,1.5,30000,1);
n=1:5;
for r=n,
    A=[A,sum(p.^r)/length(p)];
    B=[B,moment(p,r)];
end
A,B
```

运行程序，输出如下：

```
A =
    0.5043    2.5186    3.5019    19.0243    40.8658
B =
         0    2.2643   -0.0519    15.6093    -1.2970
```

由下面的语句还可以求出各阶矩的理论值。可以看出，从生成的数据求出的各阶矩和理论值的拟合程度也是很好的。

```
syms x;
A1=[];B1=[];
p1=1/(sqrt(2*pi)*1.5)*exp(-(x-0.5)^2/(2*1.5^2));
for i=1:5
    A1=[A1,vpa(int(x^i*p,x,-inf,inf),12)];
    B1=[B1,vpa(int((x-0.5)^i*p,x,-inf,inf),12)];
end
A1,B1
```

运行程序，输出如下：

```
A1=
   [0.500000000001, 2.50000000000, 3.49999999999, 1.86250000000, 40.8125000000]
B1=
   [0,          2.25000000000,          0,          15.1875000000,          0 ]
```

2. 协方差矩阵

设 $(X_1,X_2,\cdots,X_n)$ 为 n 维随机变量，称 $(E(X_1),E(X_2),\cdots,E(X_n))$ 为期望向量，称 $(D(X_1),D(X_2),\cdots,D(X_n))$ 为方差向量。

下面介绍 n 维随机变量 $(X_1,X_2,\cdots,X_n)$ 的协方差矩阵，先从二维情况讲起。

设 (X_1,X_2) 为二维随机变量，它的 4 个 2 阶中心矩存在，分别记为

$$c_{11}=E\{[X_1-E(X_1)]^2\}=D(X_1)$$
$$c_{12}=E\{[X_1-E(X_1)][X_2-E(X_2)]\}=\mathrm{Cov}(X_1,X_2)$$
$$c_{21}=E\{[X_2-E(X_2)][X_1-E(X_1)]\}=\mathrm{Cov}(X_2,X_1)=c_{12}$$
$$c_{22}=E\{[X_2-E(X_2)]^2\}=D(X_2)$$

则矩阵

$$\begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix}$$

称为随机变量(X_1,X_2)的协方差矩阵，一般地，有如下定义。

定义 3-12 设 n 维随机变量$(X_1,X_2,\cdots,X_n)$的 1+1 阶混合中心矩

$$c_{ij}=\mathrm{Cov}(X_i,X_j)=E\{[X_i-E(X_i)][X_j-E(X_j)]\},\quad i,j=1,2,\cdots$$

都存在，则称矩阵

$$\boldsymbol{C}=(c_{ij})_{n\times n}=\begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \vdots & \vdots & & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nn} \end{pmatrix}$$

为 n 维随机变量$(X_1,X_2,\cdots,X_n)$的协方差矩阵。

由于 $c_{ij}=c_{ji}$，$i,j=1,2,\cdots,n$，故协方差矩阵 $\boldsymbol{C}$ 是一个对称矩阵且主对角线元素

$$c_{ii}=D(X_i)$$

【例 3-19】 设(X_1,X_2)服从二维正态分布 $N(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)$，其联合概率密度为

$$p(x_1,x_2)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\cdot\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x_1-\mu_1)^2}{\sigma_1^2}-\frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2}+\frac{(x_2-\mu_2)^2}{\sigma_2^2}\right]\right\},$$

$$-\infty<x_1,x_2<+\infty$$

由于

$$c_{11}=D(X_1)=\sigma_1^2,\quad c_{22}=D(X_2)=\sigma_2^2,\quad c_{12}=c_{21}=\mathrm{Cov}(X_1,X_2)=\rho\sigma_1\sigma_2$$

因此协方差矩阵为

$$\boldsymbol{C}=\begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

其相应行列式为$|\boldsymbol{C}|=\sigma_1^2\sigma_2^2(1-\rho^2)$，故逆矩阵为

$$\boldsymbol{C}^{-1}=\frac{1}{|\boldsymbol{C}|}\begin{pmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{pmatrix}$$

引入向量

$$\boldsymbol{x}=\begin{pmatrix} x_1 \\ x_2 \end{pmatrix},\quad \boldsymbol{\mu}=\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}=\begin{pmatrix} E(X_1) \\ E(X_2) \end{pmatrix}\text{（期望向量）}$$

则

$$\begin{aligned}&(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{C}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\\&=\frac{1}{|\boldsymbol{C}|}(x_1-\mu_1,x_2-\mu_2)\begin{pmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{pmatrix}\begin{pmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{pmatrix}\\&=\frac{1}{1-\rho^2}\left(\frac{(x_1-\mu_1)^2}{\sigma_1^2}-\frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2}+\frac{(x_2-\mu_2)^2}{\sigma_2^2}\right)\end{aligned}$$

从而 (X_1,X_2) 的联合概率密度 $p(x_1,x_2)$ 可表示为

$$p(x_1,x_2)=\frac{1}{2\pi|\boldsymbol{C}|^{\frac{1}{2}}}\mathrm{e}^{-\frac{(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{C}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}{2}} \tag{3-34}$$

式（3-34）具有更易于推广的特点，引入向量

$$\boldsymbol{x}=\begin{pmatrix}x_1\\x_2\\\vdots\\x_n\end{pmatrix},\quad \boldsymbol{\mu}=\begin{pmatrix}\mu_1\\\mu_2\\\vdots\\\mu_n\end{pmatrix}=\begin{pmatrix}E(X_1)\\E(X_2)\\\vdots\\E(X_n)\end{pmatrix}$$

$\boldsymbol{C}$ 为协方差矩阵，则 n 维正态随机变量 $(X_1,X_2,\cdots,X_n)$ 的联合概率密度可表示为

$$p(x_1,x_2,\cdots,x_n)=\frac{1}{(2\pi)^{\frac{n}{2}}|\boldsymbol{C}|^{\frac{1}{2}}}\mathrm{e}^{-\frac{(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{C}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}{2}}$$

3.5 大数定律与中心极限定理

3.5.1 切比雪夫不等式

在实际应用中，常需要估计随机变量落在均值附近的概率，切比雪夫不等式就是解决这类问题的工具之一。

定理 3-2（切比雪夫不等式） 若随机变量的方差为实数，则

$$P\{|\xi-E(\xi)|\geqslant\varepsilon\}\leqslant\frac{D(\xi)}{\varepsilon^2},\quad\forall\varepsilon>0$$

注意： ① 该定理可以用来估算概率的界限。

② 可以在相关的概率论教科书中找到本定理的证明。

【例 3-20】 如果某大学的男生的平均身高为 175cm，标准差为 3cm，试估计身高在 166～184cm 之间的男生比例的下界。

解： 用 ξ 表示男生身高，则其均值为 $E(\xi)=175$，标准差为 $\sqrt{D(\xi)}=3$。由切比雪夫不等式，身高在 166～184cm 之间的男生比例：

$$P\{166\leqslant\xi\leqslant184\}=1-P\{|\xi-175|>9\}\geqslant1-\frac{D(\xi)}{81}=1-\frac{1}{9}=\frac{8}{9}$$

即至少有 $\frac{8}{9}\approx89\%$ 的男生的身高在 166～184cm 之间。

注意： 在本题中，如果要估计身高在 170～184cm 之间的男生比例的下界，也可以用切比雪夫不等式，但是估计的精度可能要降低。事实上：

$$\begin{aligned}P\{170\leqslant\xi\leqslant184\}&=P\{170-175\leqslant\xi-175\leqslant184-175\}\\&=P\{-5\leqslant\xi-175\leqslant9\}\geqslant P\{-5\leqslant\xi-175\leqslant5\}\end{aligned}$$

$$=1-P\{|\xi-175|>5\}\geqslant 1-\frac{9}{25}=\frac{16}{25}$$

3.5.2 大数定律

【例 3-21】 往（0, 1）区间上随机投一个质点，其坐标 $\xi\sim U(0,1)$。重复投点，将前 n 个观测值的算术平均值计算结果列在表 3-1 中，总结前 n 个观测值的算术平均值随 n 增加的变化规律。

解：从表 3-1 中可以发现：当 n 比较小时，相应的观测值的算术平均值的变化幅度比较大；随着 n 的增加，平均值的变化幅度有变化小的趋势，并且有稳定于 0.500 的趋势。因此可以猜想 n 个观测值的算术平均值随着 n 的增加而趋向于 0.5 的概率很大，即可以猜想

$$\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n}\xi_k=0.5$$

成立的概率应该很大。其中，ξ_k 表示 ξ 的第 k 次重复观测值。

表 3-1　n 个 $U(0,1)$的随机变量的算术平均值的变化规律

n	1	11	21	31	41	51	61	71
平均值	0.162	0.388	0.455	0.471	0.482	0.514	0.498	0.519
n	500	510	520	530	540	550	560	570
平均值	0.517	0.517	0.515	0.513	0.513	0.513	0.510	0.515
n	1000	1010	1020	1030	1040	1050	1060	1070
平均值	0.504	0.505	0.504	0.504	0.504	0.504	0.505	0.505
n	2000	2010	2020	2030	2040	2050	2060	2070
平均值	0.497	0.496	0.495	0.496	0.495	0.495	0.495	0.494
n	3000	3010	3020	3030	3040	3050	3060	3070
平均值	0.498	0.498	0.498	0.498	0.498	0.498	0.498	0.498
n	4000	4010	4020	4030	4040	4050	4060	4070
平均值	0.495	0.495	0.495	0.495	0.495	0.495	0.495	0.495
n	5000	5010	5020	5030	5040	5050	5060	5070
平均值	0.495	0.495	0.495	0.495	0.495	0.495	0.496	0.496
n	6000	6010	6020	6030	6040	6050	6060	6070
平均值	0.496	0.496	0.496	0.496	0.496	0.496	0.496	0.496
n	7000	7010	7020	7030	7040	7050	7060	7070
平均值	0.499	0.5000	0.499	0.499	0.499	0.499	0.499	0.499
n	8000	8010	8020	8030	8040	8050	8060	8070
平均值	0.500	0.500	0.500	0.500	0.500	0.499	0.499	0.499

其实现的 MATLAB 程序代码如下：

```
x=unidrnd(6,1000,1);
f=[];
for i=1:12
```

```
    if i<11
        n=i*10;
    elseif i==11
        n=50*10;
    else
        n=100*10;
    end
    y=x(1:n);
    f=[f;sum([y==1,y==2,y==3,y==4,y==5,y==6])/n];
end
```



运行这段程序代码后，计算出的前$i\times 10$次的各个地区结果出现的频率依次存放在 12×6维矩阵$\boldsymbol{f}$的各行中。

事实上，不仅对于均匀分布$U(0,1)$的重复观测值有上述规律，对于其他的常用分布，也有相同的规律。一般地，在概率论中有如下的定理，感兴趣的读者可以在相关的概率论教科书中找到其证明过程。

定理 3-3 若随机变量X的数学期望为有限数，则

$$\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n}X_k = E(X) \tag{3-35}$$

成立的概率为 1，这里X_k为X的第k次重复观测结果。

注意：① 考虑随机变量序列$\{X_n\}$，如果其中的每个随机变量的数学期望都是有限数，并且极限

$$\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n}[X_k - E(X_k)] = 0$$

成立的概率为 1，则称随机变量序列$\{X_n\}$满足强大数定律。

② 借助于强大数定律的概念，可以把此定理叙述为：若随机变量 X 的数学期望为有限数，X_n为X的第n次重复观测值，则$\{X_n\}$满足强大数定律。

③ 本书中把定理 3-3 称为科尔莫戈罗夫强大数定律，简称为强大数定律。

定义 3-13 假设 X 为随机变量，若X^k的数学期望存在，则称$E(X^k)$为 X 的 k 阶原点矩。

注意：k 阶原点矩是刻画随机变量的分布特征指标。例如，1 阶原点矩就是数学期望的k阶原点矩。

【例 3-22】 设随机变量X的k阶原点矩$E(X^k)$为实数，X_i为X的第i次重复观测结果。试证明样本k阶原点矩

$$\frac{1}{n}\sum_{i=1}^{n}X_i^k \xrightarrow{n\to\infty} E(X^k)$$

成立的概率是 1。

证明：记 $Y = X^k$，$Y_i = X_i^k$，$i \geqslant 1$，则 Y 的数学期望是有限数，Y_i 为 Y 的第 i 次重复观测结果。由定理 3-3 知

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}Y_i = E(Y)$$

成立的概率为 1，即

$$\frac{1}{n}\sum_{i=1}^{n}X_i^k \xrightarrow{n\to\infty} E(X^k)$$

成立的概率为 1。

【例 3-23】 考虑随机实验结果可能出现的某事件 A，重复该实验 n 次，观测到该事件出现的次数记为 $n(A)$，试证明

$$\lim_{n\to\infty}\frac{n(A)}{n} = P(A) \tag{3-36}$$

成立的概率为 1。

注意：式（3-36）解释了频率稳定于概率的原因。

证明：定义随机变量

$$X = \begin{cases} 1, & A发生 \\ 0, & A不发生 \end{cases}$$

把第 i 次实验中 X 的观测结果记为 X_i，则 $\sum_{i=1}^{n}X_i$ 表示 n 次实验中 A 出现的次数。由定理 3-3 知

$$\lim_{n\to\infty}\frac{n(A)}{n} = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}X_i = E(X) \tag{3-37}$$

成立的概率为 1。进而，X 服从两点分布，其概率密度矩阵为

$$\begin{pmatrix}0\\1\end{pmatrix}\begin{pmatrix}P(\overline{A})\\P(A)\end{pmatrix}$$

从而

$$E(X) = 0\times P(\overline{A}) + 1\times P(A) = P(A) \tag{3-38}$$

由式（3-37）和式（3-38）可知，式（3-36）成立。

定义 3-14 设 $X_1, X_2, \cdots, X_n$ 为随机变量 X 的 n 次重复观测值，称

$$F_n(x) = \frac{n(\{i \mid 1 \leqslant i \leqslant n, X_i < x\})}{n} \tag{3-39}$$

为 X 的经验分布函数，这里 $n(\{i \mid 1 \leqslant i \leqslant n, X_i < x\})$ 表示 $X_1, X_2, \cdots, X_n$ 中小于 x 的观测值的个数。

【例 3-24】 $X_1, X_2, \cdots, X_n$ 为随机变量 X 的 n 次重复观测结果，试证明对于任意给定的实数 x：

$$\lim_{n\to\infty} F_n(x) = F(x)$$

成立的概率为 1，其中 $F(x)$ 为随机变量 X 的分布函数。

证明：对于给定的实数 x，定义

$$Y = \begin{cases} 1, & X < x \\ 0, & X \geqslant x \end{cases}, \quad Y_i = \begin{cases} 1, & X_i < x \\ 0, & X_i \geqslant x \end{cases}, \quad \forall i \geqslant 1$$



则 Y 的随机变量 $Y_1, Y_2, \cdots, Y_n$ 为 Y 的 n 次重复观测结果，并且它们都是服从两点分布的离散型随机变量，有共同的概率密度矩阵

$$\begin{pmatrix} 0 \\ 1 \end{pmatrix} \begin{pmatrix} 1 - P(X < x) \\ P(X < x) \end{pmatrix}$$

由离散型随机变量数学期望的计算公式，得

$$E(Y) = P(X < x) = F(x)$$

再由定理 3-3 知

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} Y_k = E(Y) = F(x) \tag{3-40}$$

成立的概率为 1。注意到，$\sum\limits_{k=1}^{n} Y_k$ 恰好等于 n 个随机变量 $Y_1, Y_2, \cdots, Y_n$ 中等于 1 的随机变量的个数，即 $\sum\limits_{k=1}^{n} Y_k$ 恰好等于 $X_1, X_2, \cdots, X_n$ 中小于 x 的随机变量的个数，得到

$$\frac{1}{n} \sum_{k=1}^{n} Y_k = \frac{n(\{i \mid 1 \leqslant i \leqslant n, X_i < x\})}{n} \tag{3-41}$$

结合式（3-40）和式（3-41）及经验分布函数的定义，可得

$$\lim_{n\to\infty} F_n(x) = F(x)$$

成立的概率为 1。

定理 3-4 若 X 为连续型随机变量，其密度函数为 $p(x)$，$f(x)$ 是连续函数，且 $\int_{-\infty}^{+\infty} |f(x)| p(x) \mathrm{d}x < \infty$，则随机变量 $Y = f(X)$ 的数学期望

$$E(Y) = \int_{-\infty}^{+\infty} f(x) p(x) \mathrm{d}x$$

等价于

$$E[f(X)] = \int_{-\infty}^{+\infty} f(x) p(x) \mathrm{d}x$$

本定理的证明略。

【例 3-25】 设 $f(x)$ 是 $[a,b]$ 区间上非负的连续函数，试利用大数定律近似计算定积分

$$\int_{-\infty}^{+\infty} f(x) \mathrm{d}x$$

解：设 $X \sim U(a,b)$，则 X 的密度函数为

$$p(x)=\begin{cases}\dfrac{1}{b-a}, & a<x<b\\ 0, & x\leqslant a或x\geqslant b\end{cases}$$

根据定理 3-4 知

$$E[f(X)]\mathrm{d}x=\int_{-\infty}^{+\infty}f(x)p(x)\mathrm{d}x=\int_{-\infty}^{+\infty}\frac{f(x)}{b-a}\mathrm{d}x=\frac{1}{b-a}\int_{-\infty}^{+\infty}f(x)\mathrm{d}x$$

进而根据定理 3-3 知

$$\int_{-\infty}^{+\infty}f(x)\mathrm{d}x=(b-a)E[f(X)]=(b-a)\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n}f(X_k)$$

成立的概率为 1，其中 $X_1,X_2,\cdots,X_n$ 是随机变量 X 的 n 次重复观测结果。因此，当 n 充分大时，有

$$\int_{-\infty}^{+\infty}f(x)\mathrm{d}x\approx\frac{(b-a)}{n}\sum_{k=1}^{n}f(X_k)\tag{3-42}$$

注意：为利用式（3-42）近似计算定积分，需要随机变量 X 的 n 次重复观测结果，而 $X\sim U(a,b)$，因而可以通过计算机模拟产生 n 个 $U(a,b)$ 分布随机数来代替这些重复观测结果，进而可以得到积分的近似计算。这种利用随机模拟方法进行近似计算的方法也叫蒙特卡罗方法。

假设 $X\sim N(\mu,\sigma^2)$，由于理论上已经证明不能对所有的实数 x 精确地计算出正态分布函数 $\varPhi_{\mu,\sigma}(x)$ 的值，因此也就不能对所有的实数 $a<b$ 精确地计算概率

$$P(X\in[a,b])=\int_{-\infty}^{+\infty}\frac{1}{\sigma\sqrt{2\pi}}\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}\mathrm{d}x$$

式（3-42）提供了一种近似计算这个概率的方法。下面的例子演示了利用蒙特卡罗方法近似计算这种概率的具体过程。

【例 3-26】 设 $X\sim N(0,1)$，试用蒙特卡罗方法近似计算概率 $P(0.1<X<2)$。

解：在 MATLAB 命令窗口中输入代码：

```
>> Y=unifrnd(0.1,2,1,10000);
```

得到 Y 的各个分量是随机变量 $\xi\sim U(0.1,2)$ 的 10000 个重复观测的模拟结果。用 Y_k 表示 Y 的第 k 个分量，则根据式（3-39）和标准正态分布密度函数的表达式，有

$$P\{0.1<X<2\}\approx\frac{2-0.1}{10000}\sum_{k=1}^{10000}\phi(Y_k)=\frac{1.9}{10000}\sum_{k=1}^{10000}\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{Y_k^2}{2}}$$

为计算上式，在 MATLAB 中运行代码：

```
>> (1.9/(sqrt(2*pi)*10000))*sum(exp(-(Y.*Y)/2))
ans =
    0.4376
```

从计算结果可知，$P(0.1<X<2)\approx0.4376$



注意：① 在 MATLAB 中，pi 表示圆周率π的近似值。

② 函数 sqrt 用来计算平方根，如程序代码 sqrt(4)的计算结果为 2。

③ 函数 exp 用来计算以 e 为底的指数函数的值，如程序代码 exp([1,2])的计算结果为行向量$\left(e^1,e^2\right)$的数值计算结果（2.7183, 7.3891）。

④ 二元运算符号.*称为“点乘”运算符，表示两个具有相同维数的矩阵之间的一种运算，其运算结果是一个矩阵。该矩阵的维数与原来矩阵的维数相同，其任何一位置的元素等于参加运算的两个矩阵相对应位置元素的乘积。例如，[1,2].*[1,2]的运算结果为 1×2 矩阵$(1\quad 4)$，而[1,2;3,4].*[1,2;3,4]的运算结果为 2×2 矩阵$\begin{pmatrix}1 & 4\\ 9 & 16\end{pmatrix}$。

⑤ 为比较此例答案的计算精度，在 MATLAB 中运行代码：

```
>> normcdf(2,0,1)-normcdf(0.1,0,1)
ans =
     0.4374
```

比较结果得到 $P\{0.1<X<2\}$ 的计算结果是 0.4374，因此本例中的计算精度还是比较高的。

⑥ 还需要指出的是：不同的时候运行以上两句 MATLAB 代码的计算结果可能不同，这是由随机数的随机性导致的现象。事实上，对于同一问题进行两次不同的蒙特卡罗近似计算，结果一般是不同的。

3.5.3 中心极限定理

正态分布是概率论中最重要的分布。之所以说它重要，一个主要的原因是它是自然界中最常见的，而且在实际问题中经常遇到的许多随机变量都服从或近似服从正态分布。也就是说，服从正态分布的随机变量广泛存在。

那么，如何解释这种客观存在的规律性呢？本节要学习和讲解的 3 个中心极限定理，从不同的侧面给出了“什么样的随机变量及其函数服从正态分布或近似服从正态分布”。

概率论中关于论证“大量独立随机变量的和的极限分布是正态分布”的一系列定理，统称为“中心极限定理”。中心极限定理也是数理统计中关于大样本统计推断的理论依据。

定理 3-5（独立同分布-中心极限定理，即林德柏格（Lindeberg）-列维（Levy）中心极限定理） 设随机变量 $X_1,X_2,\cdots,X_n$ 相互独立，服从相同的分布，并且数学期望和方差都存在且方差不为 0，即 $E(X_i)=\mu$，$D(X_i)=\sigma^2>0$，$i=1,2,\cdots,n$，则对于任何实数 x，有下式成立：

$$\lim_{n\to\infty}P\left\{\frac{\sum_{i=1}^{n}X_i-n\mu}{\sqrt{n}\sigma}\leqslant x\right\}=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{t^2}{2}}dt$$

证明从略。下面只对本定理的结论作分析。

分析：令 $Y_n=\sum_{i=1}^{n}X_i$，则 $E(Y_n)=E\left(\sum_{i=1}^{n}X_i\right)=\sum_{i=1}^{n}E(X_i)=n\mu$

$$D(Y_n)=D\left(\sum_{i=1}^{n}X_i\right)=\sum_{i=1}^{n}D(X_i)=n\sigma^2\text{，}\ \sigma(Y_n)=\sqrt{n}\sigma$$

再设 Z_n 是 Y_n 的标准化，即

$$Z_n=\frac{Y_n-E(Y_n)}{\sigma(Y_n)}$$

那么，该定理的结果可写成

$$\lim_{n\to\infty}P\{Z_n\leqslant x\}=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{t^2}{2}}\mathrm{d}t$$

等式右边是标准正态分布的分布函数 $\Phi(x)$。这说明当 n 充分大时，Z_n 的分布将趋于标准正态分布 $N(0,1)$，而 Z_n 是 $Y_n=\sum_{i=1}^{n}X_i$ 的标准化，那么 Y_n 的分布趋于 $N(n\mu,n\sigma^2)$。也就是说，当 n 充分大时，独立同分布的随机变量 $X_1,X_2,\cdots,X_n$ 的和 $Y_n=\sum_{i=1}^{n}X_i$ 将近似服从正态分布 $N(n\mu,n\sigma^2)$。反过来讲，如果被研究的随机变量 Y 可以表示为大量独立同分布的随机变量 $X_1,X_2,\cdots,X_n$ 的和 $Y=\sum_{i=1}^{n}X_i$，其中每个随机变量 X_i 对总和 Y 只起微小的作用，那么，可以认为这个随机变量 Y 实际上是服从或近似服从正态分布 $N(n\mu,n\sigma^2)$ 的。

比如，在进行某种观测时，不可避免地会有许多随机因素影响观测结果，产生误差。有些误差是由测量仪器的精密度引起的，精密度可以在温度、大气压力或其他因素的影响下改变。有些误差是属于观测者的个人误差，大都是由观测者的视觉、听觉等引起的。这些因素中的每一个因素都可能使观测结果产生很小的误差，所有的这些误差共同影响着观测结果，于是就得到一个“总误差”。

因此，实际观测得到的误差可以看做是一个随机变量，它是许多数值微小的独立随机变量的总和。按“中心极限定理”，这个随机变量“总误差”应服从正态分布。

下面来看一个应用“独立同分布-中心极限定理”来解决实际问题的例子。

【例 3-27】 设有 30 个电子元件 $D_1,D_2,\cdots,D_{30}$，其寿命分别为 $T_1,T_2,\cdots,T_{30}$，都服从参数为 $\frac{1}{10}$ h 的指数分布，即 $T_i\sim \mathrm{e}\left(\frac{1}{10}\right)$，$i=1,2,\cdots,30$。它们的使用情况如下：$D_i$ 损坏后立即使用 D_{i+1}，$i=1,2,\cdots,29$。求这批电子元件使用的总计时间 T 不小于 350h 的概率。

解： 显然 $T_i\sim \mathrm{e}\left(\frac{1}{10}\right)$，$i=1,2,\cdots,30$，其概率密度函数为

$$p(x)=\begin{cases}\frac{1}{10}\mathrm{e}^{-\frac{1}{10}x}, & x\geqslant 0\\ 0, & \text{其他}\end{cases}$$

那么 $E(T_i)=\frac{1}{\lambda}=10$，$D(T_i)=\frac{1}{\lambda^2}=100$。

总计时间$T=\sum_{i=1}^{30}T_i$，$E(T)=\sum_{i=1}^{30}E(T_i)=300$，$D(T)=\sum_{i=1}^{30}D(T_i)=3000$。

要求的是$P\{T>350\}$，而$T>350 \Rightarrow T-E(T)>350-E(T) \Rightarrow T-300>350-300$

进而$\frac{T-E(T)}{\sqrt{D(T)}}=\frac{T-30\times10}{\sqrt{30\times100}}>\frac{350-300}{\sqrt{3000}}=0.913$



那么$P\{T>350\}=P\left\{\frac{T-E(T)}{\sqrt{D(T)}}>0.913\right\}\approx1-\Phi(0.913)=1-0.8186=0.1814$。

这说明电子元件使用的总计时间T不小于350h的概率近似等于18.14%。

下面再来讲解一个“独立同分布-中心极限定理”的特殊情形。

定理 3-6（棣莫弗（De Moivre）-拉普拉斯（Laplace）中心极限定理） 设在独立试验序列中，事件A发生的概率$P(A)=p(0<p<1)$，随机变量Y_n表示“事件A在n次独立试验中发生的次数”，则对于任何实数x有下式成立：

$$\lim_{n\to\infty}P\left\{\frac{Y_n-np}{\sqrt{np(1-p)}}\leqslant x\right\}=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{t^2}{2}}dt$$

分析：显然$Y_n\sim B(n,p)$，那么

$$E(Y_n)=np\text{，}\quad D(Y_n)=npq=np(1-p)\text{，}$$

$$\frac{Y_n-np}{\sqrt{np(1-p)}}=\frac{Y_n-E(Y_n)}{\sqrt{D(Y_n)}}=\frac{Y_n-E(Y_n)}{\sigma(Y_n)}=Y_n^*$$

是对Y_n的标准变换。标准变换后的随机变量Y_n^*近似服从标准正态分布。

所以，定理 3-6 表明：当n充分大时，服从二项分布$B(n,p)$的随机变量Y_n近似地服从参数分别为np，$np(1-p)$的正态分布$N[np,np(1-p)]$。

定理 3-7（李雅普诺夫（Lyapunov）中心极限定理） 随机变量$X_1,X_2,\cdots,X_n$相互独立，且数学期望$E(X_i)=\mu_i$，方差$D(X_i)=\sigma_i^2>0$，记$B_n=\sum_{i=1}^{n}\sigma_i^2$。如果$B_n$满足如下林德柏格条件：存在正数$\delta>0$，使得当$n\to\infty$时，有

$$\lim_{n\to\infty}\frac{1}{B_n^{2+\delta}}\sum_{i=1}^{n}E\{|X_i-\mu_i|^{2+\delta}\}=0$$

则

$$\lim_{n\to\infty}P\left\{\frac{\sum_{i=1}^{n}X_i-\sum_{i=1}^{n}\mu_i}{B_n}\leqslant x\right\}=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{t^2}{2}}dt$$

该定理说明：无论随机变量$X_1,X_2,\cdots,X_n$服从何种分布，只要相互独立，期望和方差存在且方差全不为 0，在满足林德柏格条件时，它们的和$\sum_{i=1}^{n}X_n$当n很大时，就近似地服从正

态分布。

【例 3-28】 用中心极限定理证明伯努利大数定理。

证明：设在 n 次伯努利试验中，事件 A 发生的次数n_A，$P(A)=p$。$X_i=1$ 表示"第 i 次试验中事件 A 发生"，则 $n_A=\sum_{i=1}^{n}X_i$，A 发生的频率

$$f_n(A)=\frac{n_A}{n}=\frac{1}{n}\sum_{i=1}^{n}X_i$$

显然 $X_i\sim B(1,p)$，$E(X_i)=1\times p=p$，$D(X_i)=p(1-p)$

又因为 $n_A=\sum_{i=1}^{n}X_i\sim B(1,p)$，那么 $E(n_A)=np$，$D(n_A)=np(1-p)$，$\forall\varepsilon>0$，使

$$P\{|f_n(A)-p|<\varepsilon\}=P\left\{\left|\frac{n_A}{n}-p\right|<\varepsilon\right\}=P\left\{\left|\frac{n_A-np}{\sqrt{np(1-p)}}\right|<\varepsilon\sqrt{\frac{n}{p(1-p)}}\right\}$$

$$=P\left\{-\varepsilon\sqrt{\frac{n}{p(1-p)}}<\frac{n_A-np}{\sqrt{np(1-p)}}<\varepsilon\sqrt{\frac{n}{p(1-p)}}\right\}$$

由中心极限定理知：$\lim\limits_{n\to\infty}P\{|f_n(A)-p|<\varepsilon\}=\Phi(+\infty)-\Phi(-\infty)=1$

【例 3-29】 某保险公司多年的统计资料表明，在索赔客户中因被盗而索赔的占 20%，以 X 表示在随机抽查的 100 户中因被盗向保险公司索赔的户数。

求：1）写出 X 的概率分布。2）用中心极限定理计算 $P\{14\leqslant X\leqslant 30\}$。

解：1） $X\sim B(100,0.2)$，$P\{X=k\}=C_{100}^{k}0.2^{k}0.8^{100-k}$，$k=0,1,2,\cdots,100$

2） $E(X)=np=20$，$D(X)=np(1-p)=16$

$$P\{14\leqslant X\leqslant 30\}=P\left\{\frac{14-20}{4}\leqslant\frac{X-20}{4}\leqslant\frac{30-20}{4}\right\}\approx\Phi(2.5)-\Phi(-1.5)=0.927$$

【例 3-30】 设 $X\sim P(0.5)$，其 30 次重复观测结果为 $X_1,X_2,\cdots,X_{30}$，记

$$\bar{X}\triangleq\frac{1}{30}\sum_{k=1}^{30}X_k\text{，}\quad Z\triangleq\frac{\sqrt{30}(\bar{X}-0.5)}{\sqrt{0.5}}$$

用计算机模拟 Z 的重复观测结果 1000 次，将 Z 的经验分布函数 $F_{1000}(x)$ 与 $\Phi(x)$ 在点

$$x=-3+0.5k\text{，}\quad 0\leqslant k\leqslant 12$$

的值相比较，并解释比较结果。

解：在 MATLAB 命令窗口中输入代码：

```
>> y=poissrnd(0.5,1000,30);
```

得到一个 1000×30 阶的矩阵 ***y***，该矩阵的每一行可以看做 X 的一次 30 次重复观测的模拟结果。执行代码：

```
>> xm=(mean(y,2)-0.5)*sqrt(60);
```

得到一个 1000 维的列向量 *xm*，它的每个分量都是 Z 的一次重复观测的模拟结果。运行代码：



```
>> sum([xm<-3,xm<-2.5,xm<-2,xm<-1.5,xm<-1,xm<-0.5])/1000
```

得到 $\bar{X}$ 的经验分布函数在–3, –2.5, –2, –1.5, –1, –0.5 点的值：

```
ans =
    0.9960    0.0030    0.0170    0.0610    0.1760    0.3740
```

运行代码：

```
>> sum([xm<0,xm<0.5,xm<1,xm<1.5,xm<2,xm<2.5,xm<3])/1000
```

得到 $\bar{X}$ 的经验分布函数在 0, 0.5, 1, 1.5, 2, 2.5, 3 点的值：

```
ans =
    0.4700    0.6680    0.8150    0.9160    0.9650    0.9860    0.9960
```

运行代码：

```
>> normcdf(-3:0.5:3,0,1)
```

得到分布函数 $\Phi_{0,1}(x)$ 在 $x=-3+0.5k$ 点的值，$0\leqslant k\leqslant 12$。

```
ans =
  Columns 1 through 8
    0.0013    0.0062    0.0228    0.0668    0.1587    0.3085    0.5000    0.6915
  Columns 9 through 13
  0.8413    0.9332    0.9772    0.9938    0.9987
```

将所得的经验分布函数和正态分布函数的值列入表 3–2。比较两个分布函数在相同点的值，发现它们的最大误差不超过 0.06，说明用标准正态分布函数来近似 Z 的经验分布函数的效果还是比较好的。

表 3–2　F_{1000} 与 $\Phi_{0,1}$ 的比较

x	–3.0	–2.5	–2.0	–1.5	–1.0	–0.5	0.0
$\Phi_{0,1}(x)$	0.001	0.006	0.023	0.067	0.159	0.309	0.500
$F_{1000}(x)$	0.000	0.003	0.015	0.057	0.166	0.367	0.463
x	0.5	1.0	1.5	2.0	2.5	3.0	
$\Phi_{0,1}(x)$	0.691	0.841	0.933	0.977	0.994	0.999	
$F_{1000}(x)$	0.668	0.819	0.914	0.966	0.986	0.998	

注意：如果把本例中的 30 改为 1000，根据中心极限定理，$\Phi(x)$ 近似于 Z 的经验分布函数的效果应该更好。

第 4 章　统计估计及统计特征

4.1　统计图的绘制

统计工具箱提供了具体的函数，用于绘制不同用途的统计图，主要包括以下 3 种：

- Box Plots（盒状图）：用于描述数据样本，也可以用于比较不同样本的均值。
- Distribution Plots（分布图）：显示一个或多个样本的分布。
- Scatter Plots（散度图）：用于显示一对或多对变量之间的关系。

4.1.1 盒状图

统计工具箱中绘制盒状图的函数为 boxplot。其调用格式如下：

```
boxplot(x)
boxplot(x, notch)
boxplot(x, notch, 'sym')
boxplot(x, notch, 'sym', vert)
boxplot(x, notch, 'sym', vert, whis)
```

boxplot 函数用于绘制单个样本的盒状图。其中，*x* 是分析的样本；*notch*=1，得到一个有凹口的盒状图；*notch*=0，得到一个矩形盒状图；'sym'是绘图符号；*vert*=0，得到水平的盒状图；*vert*=1，得到垂直的盒状图（默认值）。

其相关函数有：anoval，kruskalwallis。

【例 4-1】 绘制样本的盒状图。

```
>> % 产生正态分布的样本
% 样本长度
N=1024;
x1=normrnd(5,1,N,1);
x2=normrnd(6,1,N,1);
x=[x1 x2];
%参数
figure(1);
sym1='*';
notch1=1;   %凹口
boxplot(x,notch1,sym1);
figure(2);
notch2=0;   %矩形
boxplot(x,notch2);
```



```
figure(3);
vert=0;   %水平
boxplot(x,notch1,'+',vert);
```

设置不同的参数后，计算得到的盒状图分别如图 4-1～图 4-3 所示。

盒状图中包括以下图形元素：

- 盒的上下边界线分别对应样本的第 25 个和第 75 个百分点处。
- 盒中间的直线是样本的中值。如果中值不在盒的中间，表明存在倾斜度。
- 盒的凹口是样本中值置信区间的图形化表示。

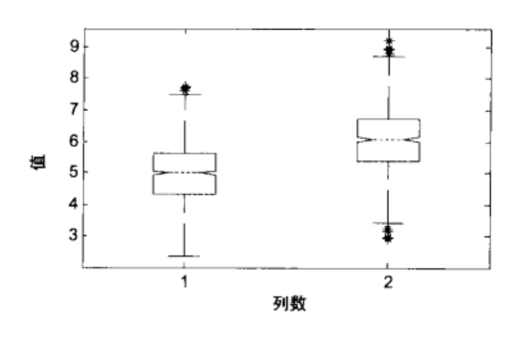


图 4-1　垂直、带凹口的盒状图

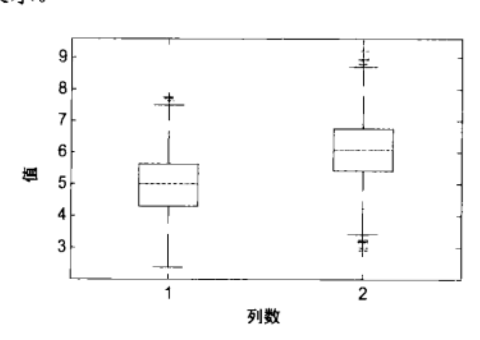


图 4-2　垂直、矩形的盒状图

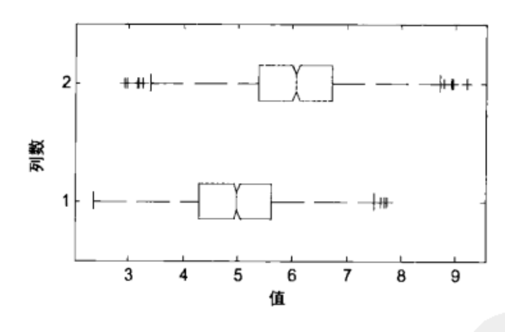


图 4-3　水平、带凹口的盒状图



4.1.2 分布图

统计工具箱提供了几种函数用于绘制一个或多个样本的分布，包括正态概率图、Quantile-Quantile 图、Weibull 概率图和累积分布图。

1. 正态概率图

绘制正态概率图的函数为 normplot。

其调用格式如下：

```
normplot(x)
h=normplot(x)
```

其中，该函数用于绘制正态概率图，用于图形化检验正态性。x 是分析的数据，当 x 是

矩阵时，对每一列显示一条直线；*h* 为返回直线的句柄。

其相关函数有：cdfplot、hist、normfit、normpdf、normrnd、normspec、normstat。

【例 4-2】 绘制正态概率图。

其实现的 MATLAB 程序代码如下：

```
%生成正态分布数据
M=100;N=1;
x=normrnd(0,1,M,N);
%生成均匀分布
y=rand(M,N);
z=[x,y];
%绘制正态概率图
h=normplot(z);
xlabel('数据');ylabel('概率');
title('正态概率图');
legend('正态分布数据','均匀分布数据');
grid on;
```

正态分布数据和均匀分布数据的概率图如图 4-4 所示。

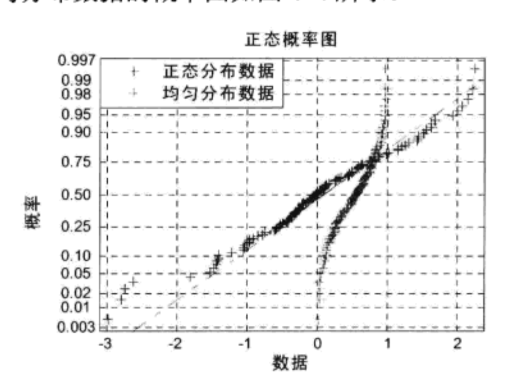


图 4-4　正态分布和均匀分布数据的概率图

在正态概率图中有 3 个图形元素："+"号表示每一个样本点数值的经验概率；实线连接了数据的第 25 个和第 75 个百分点，表示一个线性拟合；点画线将实线延伸到样本的两端。

在正态概率图中，如果所有的样本点都在实线附近，则假设样本服从正态分布是合理的；否则，如果样本不是正态分布的，则"+"号构成了一条曲线。通过观察图 4-4 中的两种不同分布样本的概率图可以验证这一点。

2．Quantile-Quantile 图

Quantile-Quantile 图可用于检验两个样本是否来自于同一分布，其函数为 qqplot。

其调用格式如下：

```
qqplot(x)
qqplot(x,y)
qqplot(x,y,pvec)
```

```
h=qqplot(...)
```

qqplot 函数用于显示一个或两个样本的 Quantile-Quantile 图。如果 x 是正态分布的，则 qqplot(x)近似于直线；如果 x，y 来自于同一分布，则 qqplot(x, y)是一条直线。x, y 是分析的样本，h 为返回直线的句柄。



其相关函数有：normplot。

【例 4-3】 绘制样本的 Quantile-Quantile 图。

```
>> %生成正态分布数据
M=100;N=1;
x=normrnd(0,1,M,N);
%生成均匀分布
y=rand(M,N);
z=[x,y];
%绘制 Quantile-Quantile 图
figure(1);
h1=qqplot(z);
xlabel('标准正态样本的 Quantile');
ylabel('输入样本的 Quantile');
title('Quantile-Quantile 图');
legend('正态分布数据','均匀分布数据');
grid on;
%生成两个正态分布样本
x=normrnd(0,1,100,1);
y=normrnd(0.5,2,50,1);
figure(2);
h2=qqplot(x,y);
xlabel('输入样本 x 的 Quantile');
ylabel('输入样本 y 的 Quantile');
title('Quantile-Quantile 图');
grid on;
%生成两个不同分布的样本
x=normrnd(5,1,100,1);
y=weibrnd(2,0.5,100,1);
figure(3);
h3=qqplot(x,y);
xlabel('输入样本 x 的 Quantile');
ylabel('输入样本 y 的 Quantile');
title('Quantile-Quantile 图');
grid on;
```

不同情况下的输出效果如图 4-5～图 4-7 所示。

3. Weibull 概率图

Weibull 概率图可用于检验一个样本是否服从 Weibull 概率分布，其函数为 weibplot。

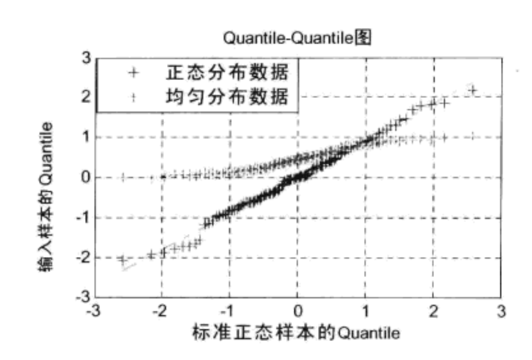


图 4-5　正态分布或均匀分布样本的 Quantile-Quantile 图

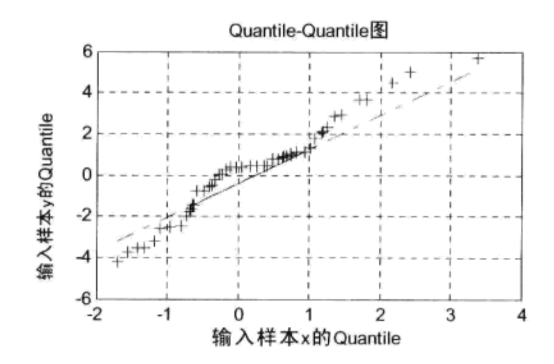


图 4-6　两个正态分布样本的 Quantile-Quantile 图

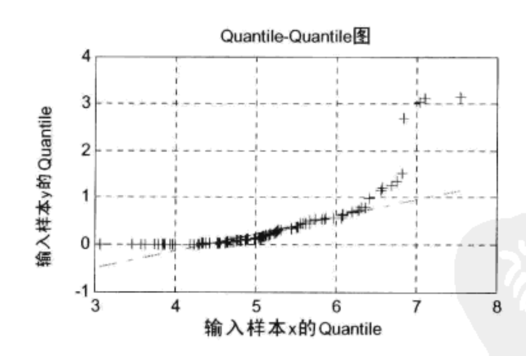


图 4-7　正态分布和 Weibull 概率分布的 Quantile-Quantile 图

其调用格式如下：

```
weibplot(x)
h=weibplot(x)
```

weibplot 函数可绘制 Weibull 概率图，用于图形化检验 Weibull 分布数据。其中，x 是分析的数据，当 x 是矩阵时，对每一列显示一条直线；h 为返回直线的句柄。

其相关函数有：weibfit、weibpdf、weibrnd。

【例 4-4】 绘制样本的 Weibull 概率图。

其实现的 MATLAB 程序代码如下：

```
>> clear all;
%生成正态分布数据
M=100;N=1;
x=normrnd(2,1,M,N);
%生成 Weibull 分布
y=weibrnd(2,0.5,100,1);
z=[x,y];
%绘制正态概率图
h=weibplot(z);
xlabel('数据');ylabel('概率');
title('Weibull 概率图');
legend('正态分布数据','Weibull 分布数据');
grid off;
```

运行程序，效果如图 4-8 所示。

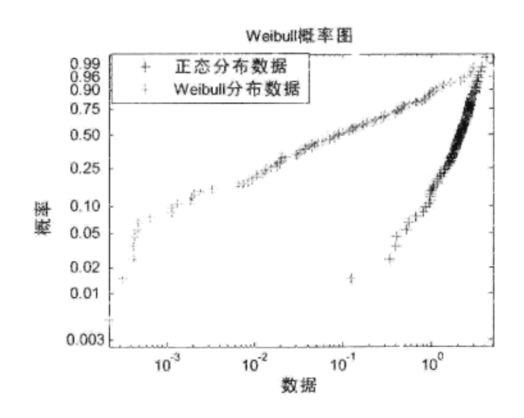


图 4-8 正态分布和 Weibull 分布数据的 Weibull 概率图

由图 4-8 可见，对正态分布样本，输出的是曲线，而对 Weibull 概率分布的样本，输出的是一条直线。

4．累积分布图

如果不想假设样本服从于一个具体的分布，则可以利用 cdfplot 函数绘制累积分布图。其调用格式如下：

```
cdfplot(x)
h=cdfplot(x)
[h,stats]=cdfplot(x)
```

cdfplot 函数用于绘制累积分布图。其中，*x* 是分析的样本；*h* 为返回曲线的句柄。

其相关函数有：ecdf、hist、kstest、kstest2、lillietest、normplot。



【例 4-5】 绘制累积分布函数图。

其实现的 MATLAB 程序代码如下：

```
>> clear all;
%生成正态分布数据
M=100;N=1;
x=normrnd(2,1,M,N);
%生成 Weibull 分布
y=weibrnd(2,0.5,100,1);
%绘制正态概率图
h1=cdfplot(x);
hold on;
h2=cdfplot(y);
xlabel('样本数据');ylabel('累积分布函数 F(x)');
title('Weibull 概率图');
legend('正态分布数据','Weibull 分布数据');
grid off;
```

运行程序，效果如图 4-9 所示。

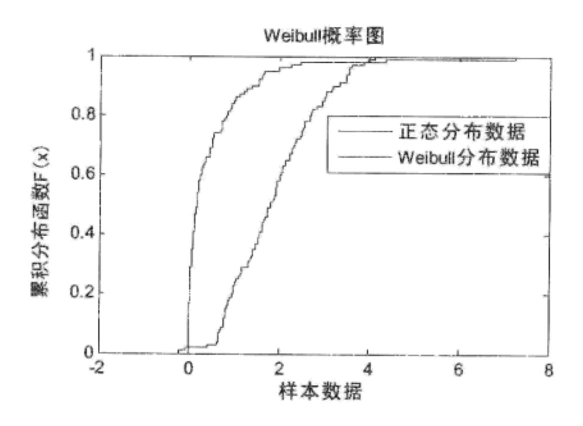


图 4-9　正态分布和 Weibull 分布样本的累积分布图

4.1.3 散度图

散度图是显示一个变量对另一个变量的一种简单图形，它可以用于确定两个变量的值或两个变量之间的关系是否属于同一组，其函数为 gscatter。其调用格式如下：

```
gscatter(x, y, group)
gscatter(x,y,group)
gscatter(x,y,group,clr,sym,siz)
gscatter(x,y,group,clr,sym,siz,doleg)
gscatter(x,y,group,clr,sym,siz,doleg,xname,yname)
h = gscatter(...)
```

gscatter 函数用于绘制不同组样本的散度图。其中，输入参数 x, y 是具有相同大小的向

量；group 是组的标记；clr，sym 是绘图的颜色和符号；siz 是大小的向量；doleg 控制是否显示图的标记；xname，yname 是 x 和 y 轴的名称。h 为返回图形中直线的句柄。

其相关函数有：gplotmatrix、grpstats、scatter。



【例 4-6】 比较 3 种不同年代汽车的重量和里程数。

其实现的 MATLAB 程序代码如下：

```
>> %装载数据
load carsmall
%比较不同类型汽车的重量和里程数
gscatter(Weight,MPG,Model_Year,'','xos');
xlabel('重量');
ylabel('里程数');
```

运行程序，效果如图 4-10 所示。

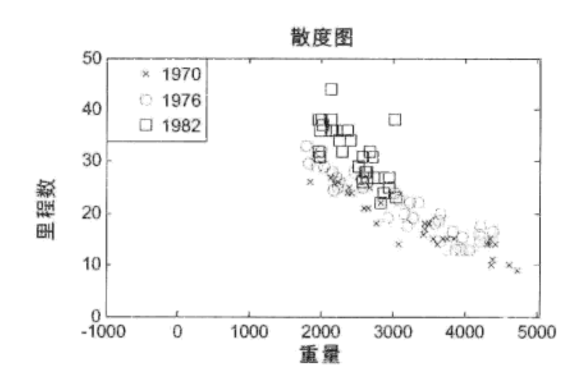


图 4-10　3 种不同年代汽车的重量和里程数的散度图

由图 4-10 可以看出，1982 年生产的汽车的重量和里程数明显区别于其他两种汽车。

4.2 变量分布估计

4.2.1 频率分布表与频率直方图

频率分布表是对连续变量的观测数据进行分组整理和初步分析的一种重要的统计数据表。频率直方图是频率分布表的图形化。通过频率分布表与频率直方图，可以实现对变量分布形态（概率密度曲线）的初步估计。掌握频率分布表的编制与频率直方图的绘制方法是统计应用的一项基本技能。

下面举例说明频率分布表的编制和频率直方图的绘制方法。

【例 4-7】 钢材中的含硅量 X 是影响材料性能的一项重要因素。在炼钢的过程中，由于各种随机因素的影响，各炉钢的含硅量 X 是有差异的。对含硅量 X 的概率分布的了解是有关钢材料性能分析的重要依据。某炼钢厂 120 炉正常生产的 25MnSi 钢的含硅量（%），即硅的质量分数如下：

0.86 0.83 0.77 0.81 0.81 0.80 0.79 0.82 0.82 0.81
0.82 0.78 0.80 0.81 0.87 0.81 0.77 0.78 0.77 0.78
0.77 0.71 0.95 0.78 0.81 0.79 0.80 0.77 0.76 0.82
0.84 0.79 0.90 0.82 0.79 0.82 0.79 0.86 0.81 0.78
0.82 0.78 0.73 0.84 0.81 0.81 0.83 0.89 0.78 0.86
0.78 0.84 0.84 0.75 0.81 0.81 0.74 0.78 0.76 0.80
0.75 0.79 0.85 0.78 0.74 0.71 0.88 0.82 0.76 0.85
0.81 0.79 0.77 0.81 0.81 0.87 0.83 0.65 0.64 0.78
0.80 0.80 0.77 0.84 0.75 0.83 0.90 0.80 0.85 0.81
0.82 0.84 0.85 0.84 0.82 0.85 0.84 0.82 0.85 0.84
0.81 0.77 0.82 0.83 0.82 0.74 0.73 0.75 0.77 0.78
0.87 0.77 0.80 0.75 0.82 0.78 0.78 0.82 0.78 0.78

下面介绍如何编制频率分布表，以及绘制频率直方图的 MATLAB 实现方法。

```
>> clear;
X=[0.86 0.83 0.77 0.81 0.81 0.80 0.79 0.82 0.82 0.81...
   0.82 0.78 0.80 0.81 0.87 0.81 0.77 0.78 0.77 0.78...
   0.77 0.71 0.95 0.78 0.81 0.79 0.80 0.77 0.76 0.82...
   0.84 0.79 0.90 0.82 0.79 0.82 0.79 0.86 0.81 0.78...
   0.82 0.78 0.73 0.84 0.81 0.81 0.83 0.89 0.78 0.86...
   0.78 0.84 0.84 0.75 0.81 0.81 0.74 0.78 0.76 0.80...
   0.75 0.79 0.85 0.78 0.74 0.71 0.88 0.82 0.76 0.85...
   0.81 0.79 0.77 0.81 0.81 0.87 0.83 0.65 0.64 0.78...
   0.80 0.80 0.77 0.84 0.75 0.83 0.90 0.80 0.85 0.81...
   0.82 0.84 0.85 0.84 0.82 0.85 0.84 0.82 0.85 0.84...
   0.81 0.77 0.82 0.83 0.82 0.74 0.73 0.75 0.77 0.78...
   0.87 0.77 0.80 0.75 0.82 0.78 0.78 0.82 0.78 0.78];
```

（1）数据分组

① 确定数据组个数。根据样本容量 n 确定分组数 k，推荐公式为 $k=1.87(n-1)^{2/5}$。

② 计算极差。计算公式为 $R_n=\max(x_1,x_2,\cdots,x_n)-\min(x_1,x_2,\cdots,x_n)$。

③ 确定组距。计算公式为 $d\approx R_n/k$，一般取 d 为数据的最小测量单位的整数倍。

④ 确定各组端点。计算公式为 $a_k=a_0+dk\ (k=0,1,\cdots,n)$，其中，$a_0<\min\{x\}$，$a_n>\max\{x\}$。$a_0$ 的确定方法：一般地，a_0 比数据的最小值小半个测量单位。

（2）统计各组频数

各组频数就是数据落入各个小组中的个数，记为 n_i。

上述计算的 MATLAB 实现由两步完成：第一步，先确定分组数的推荐公式，求出分组数 k；第二步，由 MATLAB 的 hist 函数完成计算极差、确定组距、确定各组端点和统计各组频数的工作。hist 函数的输入参数有两个，第一个是数据向量，第二个是小组个数；hist 函数的输出参数有两个，第一个输出参数返回各组的数据频数，第二个输出参数返回各个数据组的区间位置值（组中值）。

```
k=ceil(1.87*(length(X)-1)^0.4);
[ni,ak]=hist(X,k);
```



（3）计算频率

① 计算各组频率。计算公式为 $f_i = n_i/n$。

其实现的 MATLAB 程序代码如下：

```
>> fi=ni/length(X);
```

② 计算各组累积频率。计算公式为 $F_i = \sum_{j=1}^{i} f_j (i = 1,2,\cdots,k)$。

其实现的 MATLAB 程序代码如下：

```
>> mfi=cumsum(fi);
```

（4）编制频率分布表

逐一运行上述 MATLAB 程序代码，再运行如下的程序代码：

```
>> stats=[[1:k]',ak',ni',fi',mfi']
```

就可得到 120 炉的 25MnSi 钢的含硅量数据的频率分布，稍加整理后结果见表 4-1。

表 4-1　120 炉的 25MnSi 钢的含硅量数据的频率分布表

组　序	组 中 值	频　数	频　率	累 积 频 率
1	0.6519	2.0000	0.0167	0.0167
2	0.6758	0	0	0.0167
3	0.6996	2.0000	0.0167	0.0333
4	0.7235	2.0000	0.0167	0.0500
5	0.7473	8.0000	0.0667	0.1167
6	0.7712	29.0000	0.2417	0.3583
7	0.7950	15.0000	0.1250	0.4833
8	0.8188	36.0000	0.3000	0.7833
9	0.8427	15.0000	0.1250	0.9083
10	0.8665	6.0000	0.0500	0.9583
11	0.8904	4.0000	0.0333	0.9917
12	0.9142	0	0	0.9917
13	0.9381	1.0000	0.0083	1.0000

接下来介绍频率直方图和累积频率折线图及其绘制方法。

频率直方图是连续性变量频率分布的图形化，累积频率折线图是累积频率分布的图形化。

在频率直方图中，横轴表示观测变量的观测值，每一个小矩形的水平边长等于组距；纵轴表示各组数据的频率，由于频率密度曲线下方的面积恒等于 1，因此为保证直方图中所有的矩形面积之和也等于 1，规定每个小矩形的高度等于该组数据的频率/组距。

在 MATLAB 中，绘制直方图的函数是 hist 或 histfit。需要指出的是，为了便于观察，这两个函数绘制出的图形的纵轴刻度是频数值。

hist 函数在前面内容中已经见过，当有输出参数时，它将完成各组频数的统计工作；若

无输出参数，则直接绘制频率直方图。

```
>> hist(X)    %画直方图
h=findobj(gca,'Type','patch');   %为修饰图形提取指定属性对象的图形句柄 h
set(h,'FaceColor','y','EdgeColor','b');   %修饰,设置直方图的线条颜色与填充色
```

运行程序，效果如图 4-11 所示。

histfit 函数在绘制频率直方图的同时附加一条正态概率密度曲线。

```
h=histfit(X, 13);   %画附正态参考曲线的直方图，并提取图形句柄 h
set(h(1),'FaceColor','c','EdgeColor','w'); %修饰，设置直方图的线条颜色与填充色
set(h(2),'Color','r'); %修饰,设置正态参考曲线的颜色
```

运行程序，效果如图 4-12 所示。

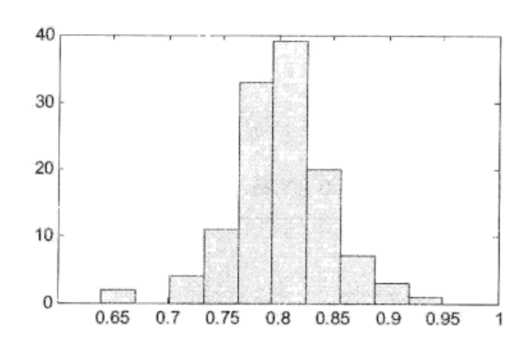


图 4-11　hist 函数绘制的直方图

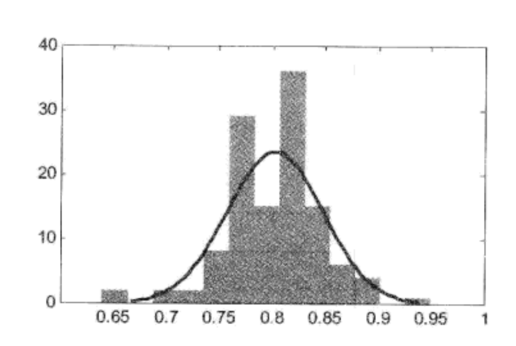


图 4-12　histfit 函数绘制的直方图

有时，人们常以频率分布表中的组中值为横坐标、以累积频率为纵坐标绘制累积频率折线（请读者用 plot 函数自行画出图形）。

在应用中，可以根据频率直方图（累积频率折线图）了解变量的概率密度曲线（分布曲线）的大致形状，进而估计变量的分布类型。在得出初步的结论后，应继续通过分布参数的估计和分布拟合检验得出更为精细的结论。

对于离散型随机变量，一般在大样条件下求样本数据的频率，画出不同数据点频率值的火柴杆图（或散点图），通过对已知的离散分布的分布律图形作出变量分布形态的估计，进一步分析参考，这里不再介绍。

下面举例说明直方图的应用。

【例 4-8】 用模拟试验的方法直观地验证定理 $\overline{X}=N(\mu,\sigma^2/n)$，$\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}}\sim N(0,1)$。

解：假设变量 $X\sim N(60,25)$，用随机数生成的方法模拟对 X 的 500 次简单随机抽样，每个样本的容量为 16。利用这 500×16 个样本数据直观地验证样本均值 $\overline{X}$ 的抽样分布为均值等于 60、方差等于 25/16 的正态分布，即 $\overline{X}\sim N(60,25/16)$。

① 用随机数生成的方法模拟简单的随机抽样。

```
>> clear;
x=[];                  %生成一个存放样本数据的空表(维数可变的动态矩阵)
```



```
for byk=1:500        %循环控制,循环执行下面的命令 500 次，在本例中相当于进行 500 次抽样
    xx=normrnd(60,5,16,1); %生成一个来自 N(60,25)、容量为 16 的样本（列向量）
    x=[x,xx];%将样本数据逐列存入列表 x,可从 MATLAB 的变量浏览器中观察这个数表
end                  %循环标志结束
```

② 计算每一个样本的样本均值，得到$\overline{x}_1,\overline{x}_2,\cdots,\overline{x}_{500}$。

```
>> xmean=mean(x);   %可从 MATLAB 的变量浏览器中观察这 500 个数据
```

③ 绘制 500 个样本均值$\overline{x}_1,\overline{x}_2,\cdots,\overline{x}_{500}$的直方图。如果直方图是单峰对称的，则可认定样本均值$\overline{X}$的抽样分布是正态分布。

```
>> k=ceil(1.87*(length(x)-1)^(2/5)); %确定分组数
h=histfit(xmean,k);   %绘制附正态参考曲线的数据 x̄1,x̄2,…,x̄500 的直方图
set(h(1),'FaceColor','c','EdgeColor','w');   %修饰,设置直方图的线条颜色与填充色
```

运行程序，效果如图 4-13 所示。

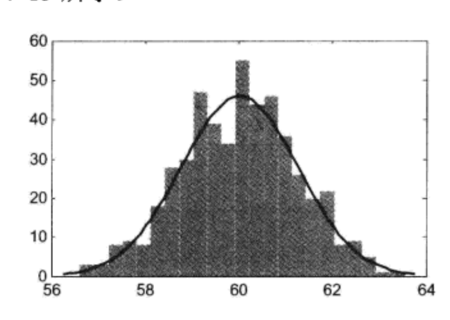


图 4-13　样本均值数据的直方图

④ 用这 500 个样本均值数据验证$\overline{X}$的均值等于 60，方差等于 25/16=1.5625。

```
>> M=mean(xmean)        %求 x̄1,x̄2,…,x̄500 的均值，以此作为 E(X̄) 的近似值
V=var(xmean)            %求 x̄1,x̄2,…,x̄500 的方差，以此作为 var(X̄) 的近似值
```

运行程序，输出如下：

```
M =    60.0133
V =     1.5649
```

上述结果表明，样本均值$\overline{X}$的抽样分布是正态的，且用$\overline{x}_1,\overline{x}_2,\cdots,\overline{x}_{500}$的样本均值与样本方差近似$\overline{X}$的数学期望与方差的效果较好。这就直观地验证了$\overline{X}=N(\mu,\sigma^2/n)$，$\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}}\sim N(0,1)$。

4.2.2 五数概括与盒状图

度量数据分布特征常用的统计量包括样本峰度、样本偏度和百分比分位数。现在先来介绍其相关概率。

（1）数据集中性的度量

数据集中性的度量见表 4-2。

表 4-2　数据集中性的度量

统计量名称	统计量的数学定义	MATLAB 函数
样本均值	$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$	mean
样本中值	$m_{0.5}$（参见样本的经验分位数）	median
样本几何均值	$\bar{x}_g=\left(\prod_{i=1}^{n}x_i\right)^{\frac{1}{n}}$	geomean
样本调和均值	$\bar{x}_h=n\left(\sum_{i=1}^{n}x_i^{-1}\right)^{-1}$	harmmean

（2）数据变异性的度量

数据变异性的度量见表 4-3。

表 4-3　数据变异性的度量

统计量名称	统计量的数学定义	MATLAB 函数
样本方差	$s^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$	var
样本标准差	$s=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$	std
样本极差	$R=x_{(n)}-x_{(1)}$	range
样本内 4 分位数的间距	$I=m_{0.75}-m_{0.25}$（参见样本的经验分位数）	iqr

（3）数据分布特征的度量

数据分布特征的度量见表 4-4。

表 4-4　数据分布特征的度量

统计量名称	统计量的数学定义	MATLAB 函数
样本的百分比分位数	$m_p=\begin{cases}x_{([np+1])}, & np\notin N\\ 0.5(x_{(np)}+x_{(np+1)}), & np\in N\end{cases}$	prctile
样本峰度	$KU=\frac{B_4}{B_2^2}$	kurtosis
样本偏度	$SK=\frac{B_3}{B_2^{\frac{3}{2}}}$	skewness

（4）两组数据线性相依程度的度量

两组数据线性相依程度的度量见表 4-5。

下面对度量数据分布特征常用统计量的几个概率作进一步说明。

样本峰度 $KU=\frac{B_4}{B_2^2}$ 是对单峰分布曲线“峰的平坦程度”或者说“曲线在其峰值附近的

陡峭程度”的度量。对于样本峰度的定义，不同文献有所不同，一般定义为 $KU=\frac{B_4}{B_2^2}-3$，此时正态分布具有零峰度。这里，采用了 MATLAB 系统中样本峰度的定义，正态分布的峰度为 3。当变量的样本峰度大于 3 时，其密度曲线比正态分布密度曲线陡峭；当变量的样本峰度小于 3 时，其密度曲线比正态分布密度曲线平坦。这里，$v_k=\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^k(k>0)$ 是样本的 k 阶中心距 B_k 的观测值。



表 4-5　两组数据线性相依程度的度量

统计量名称	统计量的数学定义	MATLAB 函数
样本协方差	$c=\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i-\overline{x})(y_i-\overline{y})$	cov
样本相关系数	$r=\frac{c}{s_x s_y}$	corrcoef

样本偏度 $SK=\frac{B_3}{B_2^{\frac{3}{2}}}$ 是对变量的分布围绕其均值的对称情况的度量。如果样本偏度等于 0，则变量分布的形状是对称的（如正态分布）；如果样本偏度大于 0，则变量分布的形状是右尾长，变量取值的密度左边偏大，称为正（或右）偏；如果样本偏度小于 0，则变量分布的形状是左尾长，变量取值的密度右边偏大，称为负（或左）偏。

样本的百分比分位数也称为样本 p 的分位数，表示如下：

$$m_p=\begin{cases} x_{([np+1])}, & np\notin N \\ 0.5(x_{(np)}+x_{(np+1)}), & np\in N \end{cases}$$

其中，N 为正整数集。关于样本的百分比分位数，应用最多的是样本的 4 分位数 $Q_1=m_{0.25}$，$Q_2=m_{0.5}$ 和 $Q_3=m_{0.75}$，分别称为第一 4 分位数、第二 4 分位数与第三 4 分位数，它反映了有 1/4 的数据小于 Q_1，有 1/4 的数据大于 Q_3，有一半的数据介于 Q_1 与 Q_3 之间。

接下来给出这几个概念在估计变量分布形态方面的一种综合应用——五数概括与 box 图。

在统计应用中，常用样本数据的最小值、最大值和 4 分位数来反映变量分布的信息，称为五数概括，而盒状图则是五数概括的图形化，如图 4-14 所示。

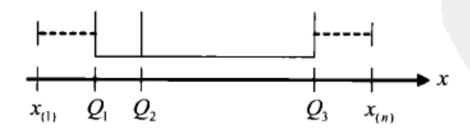


图 4-14　盒状图

从 box 图可以看出样本数据的如下特征，并可由此来推测变量的分布特点。

① 中心位置。中位数 $Q_2=m_{0.5}$ 所在的位置即为样本数据的中心，在 $[x_{(1)},Q_2]$ 和 $[Q_2,x_{(n)}]$ 中各包含一半的样本数据。

② 散布情况。样本数据全部位于 $[x_{(1)},x_{(n)}]$ 内，若将样本数据等分成 4 份，那么在区间

$[x_{(1)}, Q_1]$，$[Q_1, Q_2]$，$[Q_2, Q_3]$和$[Q_3, x_{(n)}]$内各占 1/4。各区间较短时，特别是$[x_{(1)}, x_{(n)}]$与$[Q_1, Q_3]$较短时，表示样本较集中；反之，较为分散。

③ 偏度。如果矩形位于中间位置，中位数又位于矩形的中间位置，则分布较为对称，否则是偏态分布。如果矩形偏于左端（或右端），中位数偏于矩形左端（或右端），可知分布是正偏（或负偏），此时右（左）尾较长。

④ 离群值。当矩形两端线段长度相差过大时，表明长的一侧有特大（或特小）值，称为离群值，用“+”标记，而线段终于$x_{(n-1)}$（或$x_{(2)}$），甚至终于$x_{(n-2)}$（或$x_{(3)}$）。

【例 4-9】 设有两个教学班，各有 30 名学生。在数学课程上，A 班用新教学方法组织教学，B 班用传统方法组织教学，现得期末考试成绩如下：

A:82,92,77,62,70,36,80,100,74,64,63,56,72,78,68,65,72,80,58,92,79,92,65,56,85,73,61,71,42,89

B:57,67,64,54,77,65,71,58,59,69,67,84,63,95,81,46,49,60,64,66,74,55,58,63,65,68,76,72,48,72

试在同一坐标轴上画出相应的盒状图，并对两个班的成绩进行初步的分析比较。

MATLAB 绘制盒状图的命令是 boxplot。

```
>> clear all;
X=[82,92,77,62,70,36,80,100,74,64,63,56,72,78,68,65,72,80,58,92,79,92,65,56,
85,73,61,71,42,89; 57,67,64,54,77,65,71,58, 59,69,67,84,63,95,81, 46,49,60,64,66,
74,55,58,63,65,68,76,72,48,72];
boxplot(X')  %boxplot 命令将输入矩阵的每一列视为一个变量(的样本数据)
```

运行程序，效果如图 4-15 所示。

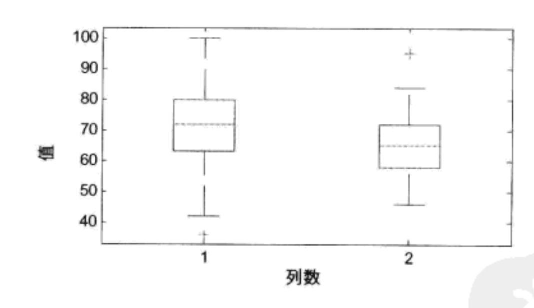


图 4-15　两个班的成绩的盒状图

从图 4-15 中可以直观地看出，两个班的数学成绩的分布是正态（对称）的，A 班成绩较为分散（方差大），B 班成绩则较为集中（方差小）。A 班成绩明显高于 B 班（均值比较，并且 A 班 25%低分段上限接近 B 班的中值线，A 班的中值线接近 B 班 25%高分段下限），A 班的平均成绩约为 70 分（中值），B 班约为 65 分（中值），A 班有一名同学的成绩过低（离群），而 B 班成绩优秀的只有一人（离群）。需要注意的是，从图 4-15 中不能得出新教学方法一定优于传统教学方法的结论，因为并不知道两个班级的学生原有的数学基础是怎样的。

4.3　参数的点估计

点估计的中心任务是通过样本构造参数的估计量，有了估计量便有了估计值。本节讲述

两个问题：一是介绍两种常用的构造统计量的方法；二是建立估计量优良性的评判标准。

设总体 X 的分布类型 $F(x;\theta)$ 已知，θ 是待估计参数。所谓参数的点估计，是指从该总体中抽取样本 $X_1,X_2,\cdots,X_n$，由样本提供的信息对未知参数作出估计。一般是建立适当的统计量 $\hat{\theta}(X_1,X_2,\cdots,X_n)$，当样本观察值为 $x_1,x_2,\cdots,x_n$ 时，以 $\hat{\theta}(x_1,x_2,\cdots,x_n)$ 作为 θ 的估计值，这种用统计量来估计总体未知参数的方法称为参数的点估计法，称 $\hat{\theta}(X_1,X_2,\cdots,X_n)$ 为 θ 的估计量。若总体中有 t 个未知参数，则要建立 t 个未知参数的估计量。在不强调估计量和估计值的区别时，通常用“估计”这个笼统的称呼。



构造估计量的方法有很多种，如矩估计法，极大似然估计法、最小二乘法、贝叶斯方法等。

4.3.1 矩估计法

由辛钦大数定律与科尔莫戈罗夫强大数定律知：如果总体 X 的 k 阶矩 $E(X^k)$ 存在，则样本 $X_1,X_2,\cdots,X_n$ 的 k 阶矩 $A_k=\frac{1}{n}\sum_{i=1}^{n}X_i^k$ 依概率收敛于总体的 k 阶矩 $E(X^k)$；样本矩的连续函数依概率收敛于总体矩的连续函数。这就启发我们可以用样本矩作为总体矩的估计量。这种用相应的样本矩去估计总体矩的估计方法称为矩估计法。

设总体的分布函数中含有 k 个未知参数 $\theta_1,\theta_2,\cdots,\theta_k$，假定总体的 k 阶矩 $E(X^k)$ 存在，则总体的 l 阶矩 $E(X^l)$ $(1\leqslant l\leqslant k)$ 是 $\theta_1,\theta_2,\cdots,\theta_k$ 的函数。用样本的 l 阶矩作为总体的 l 阶矩的估计，则得到 k 个方程（称为矩方程组）

$$\hat{a}_l(\theta_1,\theta_2,\cdots,\theta_k)=\frac{1}{n}\sum_{i=1}^{n}X_i^l,\quad l=1,2,\cdots,k$$

解此方程组，得到 $\theta_1,\theta_2,\cdots,\theta_k$ 的解 $\hat{\theta}_1(X_1,X_2,\cdots,X_n)$，…，$\hat{\theta}_k(X_1,X_2,\cdots,X_n)$，分别称 $\hat{\theta}_1(X_1,X_2,\cdots,X_n)$，…，$\hat{\theta}_k(X_1,X_2,\cdots,X_n)$ 为 $\theta_1,\theta_2,\cdots,\theta_k$ 的矩估计量。相应地，把估计量的观察值 $\hat{\theta}_1(x_1,x_2,\cdots,x_n),\cdots$，$\hat{\theta}_k(x_1,x_2,\cdots,x_n)$ 称为 $\theta_1,\theta_2,\cdots,\theta_k$ 的矩估计值。

【例 4-10】 设总体 X 的概率密度函数为

$$f(x,\theta)=\begin{cases}\theta x^{\theta-1}, & 0<x<1\\ 0, & \text{其他}\end{cases}$$

其中，$\theta>0$ 为未知参数，$X_1,X_2,\cdots,X_n$ 为来自 X 的样本，试求 θ 的矩估计。

解：因为

$$E(X)=\int_{-\infty}^{+\infty}xf(x)\mathrm{d}x=\int_0^1 x\theta x^{\theta-1}\mathrm{d}x=\int_0^1\theta x^{\theta}\mathrm{d}x=\frac{\theta}{1+\theta}$$

令 $\bar{X}=E(X)=\frac{\theta}{1+\theta}$，解得 θ 的矩估计量为 $\hat{\theta}=\frac{\bar{X}}{1-\bar{X}}$。

点估计的矩法是由皮尔逊提出的，它直观、简便，对总体数学期望和方差进行估计时不需要知道总体的分布，但是它要求总体的原点矩存在，而有些随机变量（如柯西分布）的原点矩不存在，因此就不能用此方法进行参数估计。此外，一般情况下，矩估计量不具有唯一性（如泊松分布中参数 λ 的矩估计），原因在于建立矩法方程时，选取哪些总体矩用相应样

本矩代替具有一定的随意性。它常常没有利用总体分布函数所提供的信息，因此很难保证它有优良的性质。

4.3.2 极大似然估计法

下面首先举例说明极大似然估计法的数学原理。

【例 4-11】 设有甲、乙两个布袋，甲袋中有 99 个白球和 1 个黑球，乙袋中有 1 个白球和 99 个黑球。由于某种原因不能识别哪一个是甲袋，哪一个是乙袋。问能否用统计的方法识别出来？

下面对这个问题进行数学描述与分析。

设变量 X 表示袋中的白球数，则 $X \sim \begin{pmatrix} 1 & 99 \\ p & 1-p \end{pmatrix}$，$p$ 是未知的分布参数，其取值依赖于变量 X 代表的是甲袋中的白球数，还是乙袋中的白球数。显然，当变量 X 代表的是甲袋中的白球数时，与 $p=99/100$ 是等价的；变量 X 代表的是乙袋中的白球数时，与 $p=1/100$ 是等价的。

可以通过抽样（任取一袋，从该袋中任取一球，观察其颜色）的方法来确定 $p=99/100$ 还是 $p=1/100$。

设事件 A 表示“取出的一袋为甲袋”，事件 B 表示“从袋子中取出的是白球”，则

$$P(A)=0.5\ ,\quad P(B\mid A)=99/100\ ,\quad P(B\mid \bar{A})=1/100$$

假定取出的是白球。在已知取出的是白球的条件下，判断该球来自甲袋还是乙袋的问题，可由贝叶斯公式，通过比较概率 $P(A\mid B)$ 和 $P(\bar{A}\mid B)$ 的大小来作出判断。由于在一次试验中大概率事件容易发生，因此，若 $P(A|B)>P(\bar{A}|B)$，则该球来自甲袋；若 $P(A\mid B)<P(\bar{A}\mid B)$，则该球来自乙袋。

因为

$$P(A\mid B)=\frac{P(AB)}{P(B)}=\frac{P(A)P(B\mid A)}{P(A)P(B\mid A)+P(\bar{A})P(B\mid \bar{A})}$$

$$P(\bar{A}\mid B)=\frac{P(\bar{A}B)}{P(B)}=\frac{P(\bar{A})P(B\mid \bar{A})}{P(A)P(B\mid A)+P(\bar{A})P(B\mid \bar{A})}$$

这两个式子的分母相同，分子中 $P(A)=P(\bar{A})$，故其大小取决于 $P(B\mid A)$ 和 $P(B\mid \bar{A})$ 的大小，而 $P(B\mid A)$ 和 $P(B\mid \bar{A})$ 的取值恰好等于变量 X 的分布参数 p 的两个可能的取值。这说明参数的取值同逆概率 $P(B\mid A)$ 和 $P(B\mid \bar{A})$ 之间的大小是相互决定的，即 $p=99/100$ 等价于 $P(A\mid B)>P(\bar{A}\mid B)$；$p=1/100$ 等价于 $P(A\mid B)<P(\bar{A}\mid B)$。

通过计算可知，$P(A\mid B)>P(\bar{A}\mid B)$，因此 $p=99/100$，即现在取出的是甲袋。

概括上面的思想方法，就可以得到极大似然估计法的数学原理——大概率原理：大概率事件在一次试验中容易发生。或者说，在一次试验中已经发生的事件具有较大的概率，而变量的分布参数有助于关于该变量的大概率事件的发生。

接下来讲解参数的极大似然估计的方法。

设 $X_1, X_2, \cdots, X_n \sim X$，并记变量 X 的概率分布律或概率密度函数为 $p(x;\theta_1,\theta_2,\cdots,\theta_k)$，

其中$\theta_1,\theta_2,\cdots,\theta_k$是变量$X$的$k$个未知参数。

又设对样本$(X_1,X_2,\cdots,X_n)$进行一次观测，得到样本值$(x_1,x_2,\cdots,x_n)$，这相当于n个相互独立的事件$\{X_1=x_1\},\{X_2=x_2\},\cdots,\{X_n=x_n\}$在一次试验中同时发生，即事件$\{X_1=x_1, X_2=x_2,\cdots,X_n=x_n\}$应该有较大的概率值。



（1）X是离散变量的情形

根据前述极大似然估计法的数学原理，可令

$$P\{X_1=x_1,X_2=x_2,\cdots,X_n=x_n\}=\prod_{i=1}^{n}P\{X_i=x_i\}=\prod_{i=1}^{n}P\{x_i;\theta_1,\theta_2,\cdots,\theta_k\}$$

达到最大值，此时对应的参数值$\hat{\theta}_1,\hat{\theta}_2,\cdots,\hat{\theta}_k$即为参数真值$\theta_1,\theta_2,\cdots,\theta_k$的估计值。

（2）X是连续变量的情形

对连续变量考虑概率$P\{X_1=x_1,X_2=x_2,\cdots,X_n=x_n\}$是没有意义的。因此，考虑随机点$(X_1,X_2,\cdots,X_n)$落入以点$(x_1,x_2,\cdots,x_n)$为顶点，以$\Delta x_1,\Delta x_2,\cdots,\Delta x_n$为边长的$n$维矩形区域$G$内的概率，这个概率近似等于

$$P\{(X_1,X_2,\cdots,X_n)\in G\}=\prod_{i=1}^{n}P\{x_i;\theta_1,\theta_2,\cdots,\theta_k\}\prod_{i=1}^{n}\Delta x_i$$

同理，可令这个概率达到最大值，此时对应的参数值$\hat{\theta}_1,\hat{\theta}_2,\cdots,\hat{\theta}_k$即为参数真值$\theta_1,\theta_2,\cdots,\theta_k$的估计值。

注意到，$\Delta x_i(i=1,2,\cdots,n)$与$\theta_1,\theta_2,\cdots,\theta_k$无关，使$\prod_{i=1}^{n}P\{x_i;\theta_1,\theta_2,\cdots,\theta_k\}\prod_{i=1}^{n}\Delta x_i$达到最大值的点$(\hat{\theta}_1,\hat{\theta}_2,\cdots,\hat{\theta}_k)$与使$\prod_{i=1}^{n}P\{x_i;\theta_1,\theta_2,\cdots,\theta_k\}$达到最大值的点相同，而后者在表达形式上连续型变量与离散变量是一致的，因此给出下面的定义。

定义 4-1 把样本$x_1,x_2,\cdots,x_n$的联合概率密度函数（概率分布律或概率密度函数）

$$L(\theta)=L(x_1,x_2,\cdots,x_n;\theta)=\prod_{i=1}^{n}p\{x_i;\theta\}$$

称为参数$\theta=(\theta_1,\theta_2,\cdots,\theta_k)$的似然函数。

设Θ为参数θ所有可能的取值范围，称为参数空间。若存在统计量$\hat{\theta}\in\Theta$，使得

$$L(x_1,x_2,\cdots,x_n;\hat{\theta})=\max_{\theta\in\Theta}L(x_1,x_2,\cdots,x_n;\theta)$$

则称$\hat{\theta}$为参数θ的极大似然估计量（Maximum Likelihood Estimator，MLE）。

一般情况下，求似然函数$L(\theta)$的极大值时，要先求其驻点，涉及导数运算。由于似然函数$L(\theta)$的数学表达式往往是积与幂的结构，其导数运算会比较冗繁，不方便求驻点，而对数函数$\ln x$是x的单调增函数，因此对数似然函数$\ln L(\theta)$与似然函数$L(\theta)$在同一点处取得最大值。另外，对数能够将积运算转化为和运算，将幂运算转化为积运算，从而使似然函数$L(\theta)$的数学表达式线性化，方便导数与求驻点运算。于是，通常情况下，应当先将似然函数$L(\theta)$转化为对数似然函数$\ln L(\theta)$，然后再求驻点。

【例 4-12】 求事件A发生的概率p的极大似然估计。

解：令 $X=\begin{cases}1, & \omega\in A\\ 0, & \omega\notin A\end{cases}$，其中 $\omega\in A$ 表示事件 A 发生，则 X 的概率密度函数为

$$p(x;p)=p^x(1-p)^{1-x}\quad (x=0,1)$$

故参数 p 的似然函数为

$$L(p)=\prod_{i=1}^{n}p^{x_i}(1-p)^{1-x_i}=p^{\sum\limits_{i=1}^{n}x_i}(1-p)^{\sum\limits_{i=1}^{n}(1-x_i)}$$

对数似然函数为

$$\ln L(p)=\left(\sum_{i=1}^{n}x_i\right)\ln p+\left(n-\sum_{i=1}^{n}x_i\right)\ln(1-p)$$

对 p 求导数，令导数为 0，就有

$$\frac{\mathrm{d}\ln L(p)}{\mathrm{d}p}=\frac{1}{p}\left(\sum_{i=1}^{n}x_i\right)-\frac{1}{1-p}\left(n-\sum_{i=1}^{n}x_i\right)=0$$

解得 $\ln L(p)$ 的驻点为

$$p=\frac{1}{n}\sum_{i=1}^{n}x_i$$

又在驻点处有

$$\frac{\partial^2\ln L(p)}{\partial p^2}=\frac{-n}{p(1-p)}<0$$

所以，驻点即为极大值点，即 p 的极大似然估计为 $\hat{p}=\bar{x}$。

【例 4-13】 设 $X\sim N(\mu,\sigma^2)$，求 μ 和 σ^2 的极大似然估计。

解：正态样本 $N(\mu,\sigma^2)$ 的密度函数是 $\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}$，则似然函数为

$$L(\mu,\sigma^2)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{(x_i-\mu)^2}{2\sigma^2}}=\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}}\mathrm{e}^{-\frac{\sum\limits_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}}$$

将其取对数，并令关于 μ,σ^2 的一阶导数为零，则得

$$\frac{\partial\ln L(\mu,\sigma^2)}{\partial\mu}=\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)=0$$

$$\frac{\partial\ln L(\mu,\sigma^2)}{\partial\sigma^2}=-\frac{n}{2\sigma^2}+\frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i-\mu)^2=0$$

解此关于 μ,σ^2 的方程组，得驻点

$$\mu=\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i\text{，}\quad \sigma^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$$

又可求得对数似然函数的二阶导函数矩阵是非正定矩阵，因此驻点即为似然函数的极大值点，并将 μ 的样本表达式代入 σ^2 的驻点表达式，得 μ 与 σ^2 的极大似然估计为

$$\hat{\mu} = \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i\ ,\quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

【例 4-14】 随机产生 100 个服从正态分布 $N(2,0.5^2)$ 的样本数据 X，并用这些数据估计总体 $N(\mu,\sigma^2)$ 中的参数 μ,σ，求出参数的最大似然估计值和置信水平为 99%的置信区间。



分析：随机产生的 100 个数据可视为从总体中抽出的容量为 100 的样本，样本的观测值就是这 100 个数据，可用命令 normfit(X, alpha)求出参数 μ,σ 的估计。

其实现的 MATLAB 程序代码如下：

```
>> clear all;
X=normrnd(2,0.5,100,1);   %产生 100 个样本数据
[muhat,sigmahat,muci,sigmaci]=normfit(X,0.01)
```

运行程序，输出如下：

```
muhat =     2.0240
sigmahat =     0.4343
muci =
    1.9099
    2.1380
sigmaci =
    0.3665
    0.5298
```

注意：参数 μ,σ 的估计最大似然值分别为 2.0240、0.4343，参数 μ,σ 的置信水平为 99%的置信区间分别为[1.9099, 2.1380]、[0.3665, 0.5298]。这一估计结果和总体 $N(\mu,\sigma^2)$ 中的参数真实数值 $\mu=2$，$\sigma=0.5$ 是非常接近的。

根据前面几个例子的讲解，可以概括出求极大似然估计值的一般步骤：

1）明确变量的分布律和密度函数。

2）写出似然函数 $L(\theta)$。

3）求似然函数 $L(\theta)$ 的最大值点，得 $\hat{\theta}_{MLE}$。

4）应用问题中，将样本数据代入 $\hat{\theta}_{MLE}$ 求出具体的估计值。

值得注意的是，求解对数似然方程组是建立在其可导并且导数变号的基础上的，如例 4-12 和例 4-13。若不满足这一条件，需针对似然函数 $L(\theta_1,\theta_2,\cdots,\theta_k)$ 的单调性，利用极大似然估计的基本原理直接对 $L(\theta_1,\theta_2,\cdots,\theta_k)$ 的最大值问题进行讲解。

极大似然估计量有一个简单而有用的性质：设 θ 的函数 $g=g(\theta)$ 是 Θ 上的实值函数，且有唯一反函数。如果 $\hat{\theta}$ 是 θ 的极大似然估计量，则 $g(\hat{\theta})$ 也是 $g(\theta)$ 的极大似然估计量。这个性质称为极大似然估计的不变性。根据这一性质，可以使一些复杂结构的参数的极大似然估计问题简单化。

极大似然估计法是在变量分布类型已知的情况下使用的一种参数估计法。一般地，用极大似然估计法所得的估计的性质比用矩估计法所得的性质要好，故通常多用极大似然估计法。

MATLAB 进行极大似然估计的函数为 mle。

其调用格式如下：

```
[phat, pci]=mle(data, 'distribution', dist, 'alpha', a, 'ntrials', n)
```

其中，输出参数 phat 是指定分布的参数的极大似然估计值（多参数时为行向量），pci 是参数的区间估计的置信上限和下限（与参数对应的二维列向量，可以省略）。输入参数 data 是样本数据向量（不可省略）。引用参数'distribution'及其取值 dist 设置变量的分布类型（应用中，dist 要用具体的分布名称字符串替换，并用单引号引起），二者要成对出现（可以同时默认为正态分布）。引用参数'alpha'及其取值 a 设置区间估计的显著性水平，二者要成对出现（可以同时省略，默认值为 0.05，即置信水平为 0.95）。引用参数'ntrials'及其取值 n 仅在分布类型为二项分布时引用（对于其他分布可以省略），用于设置二项分布中试验的次数。

dist 的取值包括 Beta, Bernoulli, Binomial, Discrete Uniform, Exponential, Extreme Value, Gamma, Geometric, Lognormal, Negative Binomial, Normal, Poisson, Rayleigh, Uniform, Weibull。

【例 4-15】 通常情况下，引用常数的测定值服从均值为 μ、标准差为 σ 的正态分布。某人在实验中使用金球测定引力常数，6 次测定的观察值为 6.683, 6.681, 6.676, 6.678, 6.679, 6.672。试用极大似然估计法对未知参数 μ 和 σ 作出估计。

其实现的 MATLAB 程序代码如下：

```
>> clear all;
x=[6.683,6.681,6.676,6.678,6.679,6.672];
phat=mle(x,'distribution','norm','alpha',0.05)
```

运行程序，输出如下：

```
phat =
    6.6782    0.0035
```

即 μ 的估计值为 6.6782，σ 的估计值为 0.0035。其实，在此例计算中，mle 函数的调用可以简化为 p=mle(x)。

4.3.3 估计量的性能分析

在分析和评论估计量性能的时候，常用的准则包括无偏性准则、均方误差准则和相合性准则。

1．无偏性准则

估计量是随机变量，对于不同的样本值会得到不同的估计值。用户希望估计值在未知参数的真值附近摆动，而它的期望值等于未知参数的真值。这就产生了无偏性准则。

定义 4-2（无偏估计） 设 $\hat{\theta}(X_1,X_2,\cdots,X_n)$ 是变量 X 的未知的一维参数 θ 的估计量，若 $E(\hat{\theta})=\theta$，则称 $\hat{\theta}$ 为 θ 的无偏估计，否则称为有偏估计。

定义 4-3（渐近无偏估计） 设 $\hat{\theta}(X_1,X_2,\cdots,X_n)$ 是变量 X 的未知的一维参数 θ 的有偏估计量，但是 $\lim\limits_{n\to\infty} E(\hat{\theta})=\theta$，则称 $\hat{\theta}$ 为 θ 的渐近无偏估计。

下面，不加证明地列举出关于无偏性的几个重要结论。

1）无论变量 X 服从何种分布，样本的 k 阶原点矩 $A_k=\frac{1}{n}\sum_{i=1}^{n}X_i^k(i=1,2,\cdots,n)$ 是变量 X 的 k 阶原点矩 $E(X^k)$ 的无偏估计。自然，$\bar{X}$ 是 $E(X)$ 的无偏估计。



2）无论变量 X 服从何种分布，样本（修正）方差 $S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$ 是变量 X 的方差 σ^2 的无偏估计。

3）样本方差（二阶中心矩）B_2 不是变量的方差 σ^2 的无偏估计，但是 $\lim_{n\to\infty}E(B_2)=\sigma^2$，所以 B_2 是 σ^2 的渐近无偏估计。

4）样本标准差 $S=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2}$ 不是变量 X 的标准差 σ 的无偏估计，但是，在变量的正态性假设下，可将样本标准差修正为 $\hat{\sigma}_S=C_nS$，$\hat{\sigma}_S$ 是 σ 的无偏估计，其中 $C_n=\sqrt{\frac{n-1}{2}}\frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}$ 称为正态标准差的无偏系数。由于 $\lim_{n\to\infty}C_n=1$，所以 S 是 σ 的渐近无偏估计。

无偏性准则是对估计量的一个相互要求。无偏性估计的统计意义是指估计量不产生系统性的偏差。例如，用样本均值 $\bar{X}$ 作为变量均值 μ 的估计时，由于 $\bar{X}$ 是随机变量，故在一次估计中 μ 的实现值与其真值之间存在偏差 $\bar{X}-\mu$。这种偏差是随机的，虽无法说明一次估计所产生的偏差，但是对同一统计问题大量重复使用 $\bar{X}$ 估计 μ 时，实际产生的偏差 $\bar{X}-\mu$ 随机地在 0 的周围波动，不会产生系统的 $\bar{X}$ 偏大于（小于）μ 的情况。

渐近无偏是指估计量存在系统性的偏差，但是这种系统性偏差随着样本容量的增加而趋向于消失。

2．均方误差准则

如果在样本容量 n 相同的情况下，$\hat{\theta}_1$ 的观察值较 $\hat{\theta}_2$ 的观察值更集中在真值 θ 的附近，则认为用 $\hat{\theta}_1$ 对 θ 进行的估计优于用 $\hat{\theta}_2$ 对 θ 进行的估计。

定义 4-4（均方误差准则） 设 $\hat{\theta}=\hat{\theta}(X_1,X_2,\cdots,X_n)$ 是变量 X 的未知的一维参数 θ 的估计量，称 $MSE\hat{\theta}=E(\hat{\theta}-\theta)^2$ 为估计量 $\hat{\theta}$ 的均方误差。对于参数 θ 的任意两个估计量 $\hat{\theta}_1$ 和 $\hat{\theta}_2$，若 $MSE\hat{\theta}_1\leqslant MSE\hat{\theta}_2$，且在参数空间中至少有一个 θ_0，使不等式中的“$<$”严格成立，则称在均方误差意义下 $\hat{\theta}_1$ 是优于 $\hat{\theta}_2$ 的估计。

定理 4-1（均方误差的分解定理） $MSE\hat{\theta}=Var(\hat{\theta})+[E(\hat{\theta})-\theta]^2$

事实上

$$
\begin{aligned}
MSE\hat{\theta}&=E[(\hat{\theta}-\theta)^2]=E[(\hat{\theta}-E(\hat{\theta})+E(\hat{\theta})-\theta)^2]\\
&=E[\hat{\theta}-E(\hat{\theta})^2]+2E[(\hat{\theta}-E(\hat{\theta}))][E(\hat{\theta})-\theta)]+[E(\hat{\theta})-\theta]^2
\end{aligned}
$$

由于

$$E[(\hat{\theta}-E(\hat{\theta}))(E(\hat{\theta})-\theta)]=0$$

所以

$$MSE\hat{\theta}=Var(\hat{\theta})+[E(\hat{\theta})-\theta]^2$$

若 $\hat{\theta}$ 是 θ 的无偏估计，则 $MSE\hat{\theta}=Var(\hat{\theta})$。

一个参数往往有不止一个无偏估计。由均方误差的分解定理不难理解，无偏估计以方差小者为好。

定义 4-5（最小方差无偏估计） 设 $\hat{\theta}^*(X_1,X_2,\cdots,X_n)$ 是变量 X 的未知参数 θ 的一个估计量，若 $\hat{\theta}$ 满足：

1） $E(\hat{\theta}^*)=\theta$，即 $\hat{\theta}^*$ 为 θ 的无偏估计。

2） $Var(\hat{\theta}^*)\leqslant Var(\hat{\theta})$，$\hat{\theta}(X_1,X_2,\cdots,X_n)$ 是 θ 的任意一个无偏估计，则称 $\hat{\theta}^*$ 为 θ 的最小方差无偏估计（也称最佳无偏估计）。

请注意下面几个关于最小方差无偏估计的结论：

1）最小方差无偏估计可能存在，也可能不存在。

2）对于正态变量 X，样本均值 $\bar{X}$ 和样本方差 S^2 是 μ 和 σ^2 的最小方差无偏估计。

3）极大似然估计往往是均方误差最小的估计。

均方误差准则是最为常用的估计量性能评价准则，可以这样理解它的统计意义：设 $\hat{\theta}$ 为 θ 的一个估计，由于估计量是随机变量，故在一次估计中 θ 的实现值与其真值之间存在偏差 $\hat{\theta}-\theta$。一般希望这种偏差尽可能小，但是由于偏差是随机变量，因此，不能根据一次估计时偏差 $\hat{\theta}-\theta$ 的大小来判断估计的优劣，而应根据对同一个参数 θ 用同一个估计量 $\hat{\theta}$ 进行的多次估计的“平均偏差”来判断。为避免求平均偏差时 $\hat{\theta}-\theta$ 的正负值相互抵消，使用 $(\hat{\theta}-\theta)^2$ 表示一次估计中的（平方）误差。于是，$MSE\hat{\theta}_1\leqslant MSE\hat{\theta}_2$ 表明多次用估计 $\hat{\theta}_1$ 和 $\hat{\theta}_2$ 去估计 θ，$\hat{\theta}_1$ 的观察值较 $\hat{\theta}_2$ 的观察值更密集在真值 θ 的附近。换句话说，均方误差准则说明，当使用不同的估计量 $\hat{\theta}_1$ 和 $\hat{\theta}_2$ 去估计 θ 时，其均方误差越小，估计的效果越好；反之，均方误差越大，估计的效果越差。

3. 相合性准则

无偏性准则和均方误差准则是在样本容量 n 固定的情形下讲解估计量优劣的。设变量 $X\sim F(x)$，$\hat{F}_n(x)$ 为样本的经验分布函数，由 Γ_{JIHBEHKO} 定理

$$P\left\{\lim_{n\to\infty}\sup_{-\infty<x<+\infty}\left|\hat{F}_n(x)-F(x)\right|=0\right\}=1$$

当样本容量 n 趋向于无穷时，样本的经验分布函数以概率 1 一致收敛于变量的分布函数。也就是说，当样本容量 n 趋向于无穷时，样本中包含的关于变量分布的信息不断增加，以致充分到可以将变量分布刻画到任意精确的程度。因此，有理由要求一个“好的”估计量，当样本容量 n 趋向于无穷时，在一定的数学意义下收敛于被估参数。

定义 4-6（相合估计） 设 $\hat{\theta}(X_1,X_2,\cdots,X_n)$ 为参数 θ 的估计量，若对任意的 $\varepsilon>0$，有

$$\lim_{n\to\infty}P\left\{\left|\hat{\theta}-\theta\right|\geqslant\varepsilon\right\}=0$$

而且这对θ的一切可能取的值都成立，则称$\hat{\theta}$是参数θ的一个相合估计。

相合性准则是对一个估计量最基本的要求。它说明，随着样本容量的增大，一个“好的”估计量$\hat{\theta}$应该越来越靠近参数θ的真值，使绝对偏差$|\hat{\theta}-\theta|$较大的概率越来越小。如果一个估计量没有相合性，那么不论样本取多大，也不可能把未知参数估计到预定的精度。这种估计量显然是不可取的。



下面，不加证明地列举出关于相合估计的几个重要结论。

1）相合估计具有不变性。当$\hat{\theta}_1,\hat{\theta}_2,\cdots,\hat{\theta}_k$分别是$\theta_1,\theta_2,\cdots,\theta_k$的相合估计时，若$g(\theta_1,\theta_2,\cdots,\theta_k)$为连续函数，则$g(\hat{\theta}_1,\hat{\theta}_2,\cdots,\hat{\theta}_k)$是$g(\theta_1,\theta_2,\cdots,\theta_k)$的相合估计。

2）样本的k阶原点矩$A_k=\frac{1}{n}\sum_{i=1}^{n}X_i^k$是变量$X$的$k$阶原点矩$E(X^k)$的相合估计，故样本均值$\bar{X}$是变量均值$\mu$的相合估计。

3）样本的二阶中心矩$B_2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2$是变量$X$的方差$\sigma^2$的相合估计。

4）样本方差$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$是变量的方差$\sigma^2$的相合估计，样本标准差$S=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2}$是变量的标准差$\sigma$的相合估计。

5）事件发生的频率是其概率的相合估计。

6）极大似然估计量往往具有相合性。

4.4 区间估计

点估计给出了总体参数θ的估计值$\hat{\theta}(x_1,x_2,\cdots,x_n)$，虽然简单明确，但由于它是$\theta$的一个近似值，所以与$\theta$总有偏差。在点估计中既没有反映近似值的精确度，又不知道它的偏差范围，这是点估计的缺陷。因此需要寻求另一种方法，希望这种方法能估计出一个范围，并知道这个范围包含参数真值的可信度。这种形式的估计称为区间估计。

4.4.1 区间估计的概念

1. 区间估计的含义

区间估计就是根据样本来确定统计量$\underline{\theta}(X_1,X_2,\cdots,X_n)$和$\overline{\theta}(X_1,X_2,\cdots,X_n)$，使

$$P\{\underline{\theta}(X_1,X_2,\cdots,X_n)<\theta<\overline{\theta}(X_1,X_2,\cdots,X_n)\}=1-\alpha \qquad (4\text{-}1)$$

其中，$(\underline{\theta},\overline{\theta})$为$\theta$的置信区间，$1-\alpha$为此置信区间的置信水平，$\underline{\theta}$和$\overline{\theta}$分别为置信下限和置信上限。

显然，置信区间是一个随机区间，式（4-1）的含义是：若反复抽样多次（每次取样本容量都是n），在每次取样下，对样本的观察值$x_1,x_2,\cdots,x_n$，就得到一个区间（$\underline{\theta}(X_1,X_2,\cdots,X_n)$，$\overline{\theta}(X_1,X_2,\cdots,X_n)$），每个这样的区间要么包含$\theta$的真值，要么不包含$\theta$的

真值。按伯努利大数定理，在这样多的区间中，大约有$100(1-\alpha)\%$的区间包含未知参数θ，而不包含θ的区间约占$100\alpha\%$。例如，若$\alpha=0.01$，反复抽样1000次，则得到的1000个区间中不包含θ真值的约有10个。通常α给得较小，这样式（4-1）的概率就较大。因此，置信区间的长度的平均$E(\bar{\theta}-\underline{\theta})$表达了区间估计的精确性；置信水平$1-\alpha$表达了区间估计的可靠性，它是区间估计的可靠概率，而显著性水平α表达了区间估计的不可靠概率。

置信水平$1-\alpha$一般要根据具体问题的要求来选定，并要注意：α越小，$1-\alpha$越大，即区间$(\bar{\theta}-\underline{\theta})$包含$\theta$真值的可信度越大，但区间也越长，即估计的精确度越差；反之，提高估计的精确度则会增大误判风险α，即$(\bar{\theta}-\underline{\theta})$不包含$\theta$真值的概率会增大。从后面推出的置信区间公式可看出，若其他条件不变，增大样本容量n，可以缩短置信区间的长度，从而提高精度，但增大样本容量往往不现实。因此，通常是根据不同类型的问题，先确定一个较大的置信水平$1-\alpha$，在这一前提下，寻找精度尽可能高的区间估计。如果对$\alpha=0.05$，有

$$P\left\{-1.96<\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}<1.96\right\}=0.95,\quad P\left\{-1.75<\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}<2.33\right\}=0.95$$

比较两个置信区间$\left(\bar{X}-\frac{\sigma}{\sqrt{n}}u_{0.025},\bar{X}+\frac{\sigma}{\sqrt{n}}u_{0.025}\right)$和$\left(\bar{X}-\frac{\sigma}{\sqrt{n}}u_{0.01},\bar{X}+\frac{\sigma}{\sqrt{n}}u_{0.04}\right)$，前者的区间长度$2u_{0.025}\frac{\sigma}{\sqrt{n}}=3.92\frac{\sigma}{\sqrt{n}}$比后者的区间长度$(u_{0.04}+u_{0.01})\frac{\sigma}{\sqrt{n}}=4.08\frac{\sigma}{\sqrt{n}}$短，置信区间越短表示估计的精度越高。由经验知，当$n$固定时，在给定的$1-\alpha$下，对称区间的长度最短。

2. 基本思想

对于给定值$\alpha(0<\alpha<1)$，为得到满足$P\{\bar{\theta}<\theta<\underline{\theta}\}=1-\alpha$的统计量$\underline{\theta}(X_1,X_2,\cdots,X_n)$和$\bar{\theta}(X_1,X_2,\cdots,X_n)$，将随机区间$(\underline{\theta},\bar{\theta})$包含$\theta$的概率$P\{\bar{\theta}<\theta<\underline{\theta}\}=1-\alpha$，转化成某随机变量$W(X_1,X_2,\cdots,X_n;\theta)$落在区间$(a,b)$上的概率

$$P\{a<W(X_1,X_2,\cdots,X_n;\theta)<b\}=1-\alpha$$

然后通过解不等式$a<W(X_1,X_2,\cdots X_n;\theta)<b$得到

$$\underline{\theta}(X_1,X_2,\cdots,X_n)<\theta<\bar{\theta}(X_1,X_2,\cdots,X_n)$$

为实现这个目的，所要找的函数$W(X_1,X_2,\cdots,X_n;\theta)$必须满足两个条件：

1）仅是样本$X_1,X_2,\cdots,X_n$和待估计参数θ的函数，而不再含有其他未知参数。

2）(a,b)必须是确定的。为此，要求$W(X_1,X_2,\cdots,X_n;\theta)$的分布已知。

3. 其他方法

按上述分析思路，归纳出求未知参数θ的置信区间的一般步骤如下：

1）选择一个函数$W(X_1,X_2,\cdots,X_n;\theta)$，它仅是样本$(X_1,X_2,\cdots,X_n)$和$\theta$的函数，而不再含有其他未知参数，且其分布已知（称$W(X_1,X_2,\cdots,X_n;\theta)$为统计量）。

2）对给定的置信水平$1-\alpha$，确定常数a,b，使得

$$P\{a<W(X_1,X_2,\cdots,X_n;\theta)<b\}=1-\alpha$$

3）由不等式$a<W(X_1,X_2,\cdots,X_n;\theta)<b$得到的等价不等式



$$\underline{\theta}(X_1,X_2,\cdots,X_n)<\theta<\overline{\theta}(X_1,X_2,\cdots,X_n)$$

其中，$\underline{\theta}(X_1,X_2,\cdots,X_n)$ 和 $\overline{\theta}(X_1,X_2,\cdots,X_n)$ 都是统计量，那么 $(\underline{\theta},\overline{\theta})$ 就是 θ 的置信水平 $1-\alpha$ 的置信区间。

【例 4-16】 从一批灯泡中随机抽取 5 只做寿命试验，测得寿命（单位：h）如下：1050, 1100, 1120, 1250, 1280。设灯泡寿命服从正态分布。试在 0.95 置信水平下估计灯泡的平均寿命。

分析：设 X 表示灯泡寿命，依题意知 $X\sim N(\mu,\sigma^2)$，则灯泡的平均寿命为 $E(X)=\mu$。因此本题的实质是估计正态分布参数 μ，但方差 σ^2 未知。于是，参数 μ 的估计量选用样本均值 $\overline{X}$，统计量选用 $T=\dfrac{\overline{X}-\mu}{S/\sqrt{n}}\sim t(n-1)$。对寿命问题，通常只关心寿命下限，故相应的下侧区间估计的准则为 $P\{\mu\geqslant\hat{\mu}_L\}\geqslant 1-\alpha$，其中置信下限 $\hat{\mu}_L=\overline{X}-t_{1-\alpha}(n-1)\dfrac{S}{\sqrt{n}}$（注意：单侧估计时，显著性水平 α 不再等分配置在双侧尾部，而是全部置于所关注的一侧）。

其实现的 MATLAB 程序代码如下：

```
>> clear all;
x=[1050,1100,1120,1250,1280];
N=length(x);
muEST=mean(x)
muLOWER=muEST-tinv(0.95,N-1)*sqrt(var(x)/N)
```

运行程序，输出如下：

```
muEST =          1160
muLOWER =    1.0649e+003
```

计算结果表明，这批灯泡的平均寿命约为 1160h，以 0.95 的概率保证这批灯泡的平均寿命不低于 1065h。

【例 4-17】 引力常数的测定值 $X\sim N(\mu,\sigma^2)$，今分别使用金球和铂球进行实验测定。

1）用金球测定，观察值为 6.683, 6.681, 6.676, 6.678, 6.679, 6.672。

2）用铂球测定，观察值为 6.661, 6.661, 6.667, 6.667, 6.664。

试针对 1）、2）两种情况分别对引力常数测定值的均值和标准差进行估计（置信水平为 0.9）。

分析：此问题可依正态变量分布参数的小样本估计方法，对测定值均值的估计选估计量 $\overline{X}$ 和统计量 $T=\dfrac{\overline{X}-\mu}{S/\sqrt{n}}\sim t(n-1)$，置信区间为

$$\left[\overline{X}-t_{1-\alpha/2}(n-1)\frac{S}{\sqrt{n}},\overline{X}+t_{1-\alpha/2}(n-1)\frac{S}{\sqrt{n}}\right]$$

对测定值标准差的估计选估计量 S^2 和枢轴量 $\chi^2=\dfrac{(n-1)S^2}{\sigma^2}\sim\chi^2(n-1)$，置信区间为

$$\left[\frac{n-1}{\chi^2_{1-\alpha/2}(n-1)}S^2, \frac{n-1}{\chi^2_{\alpha/2}(n-1)}S^2\right]$$

然后，依上述算法组织 MATLAB 命令进行数据处理，这里用 mle 函数进行数据处理。

其实现的 MATLAB 程序代码如下：

```
>> clear all;
x=[6.683 6.681 6.676 6.678 6.679 6.672];
y=[6.661 6.661 6.667 6.667 6.664];
[phat,pci]=mle(x,'alpha',0.1)   %金球测定的估计
[PHAT,PCI]=mle(y,'alpha',0.1)   %铂球测定的估计
```

运行程序，输出如下：

```
phat =
    6.6782    0.0035
pci =
    6.6750    0.0026
    6.6813    0.0081
PHAT =
    6.6640    0.0027
PCI =
    6.6611    0.0019
    6.6669    0.0071
```

计算结果表明，金球测定的 μ 的估计值为 6.6782，置信区间为[6.6750, 6.6813]；σ 的估计值为 0.0035，置信区间为[0.0026, 0.0081]。铂球测定的 μ 的估计值为 6.6640，置信区间为[6.6611, 6.6669]；σ 的估计值为 0.0027，置信区间为[0.0019, 0.0071]。

4.4.2 单正态总体参数的区间估计

设总体 $X \sim N(\mu,\sigma^2)$，$X_1,X_2,\cdots,X_n$ 是 X 的样本，$1-\alpha$ 为给定的置信水平，下面来确定总体均值 μ 和总体方差 σ^2 的置信区间。

1. 单正态总体均值的区间估计

（1）σ^2 已知时，均值 μ 的置信区间

以样本均值 $\overline{X}$ 作为 μ 的一个点估计，由正态公式知

$$U=\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

由正态分布的分位点知

$$P\left\{|U|<u_{\frac{\alpha}{2}}\right\}=1-\alpha$$

即

$$P\left\{\left|\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}\right|<u_{\frac{\alpha}{2}}\right\}=1-\alpha$$



或

$$P\left\{\overline{X}-\frac{\sigma}{\sqrt{n}}u_{\frac{\alpha}{2}}<\mu<\overline{X}+\frac{\sigma}{\sqrt{n}}u_{\frac{\alpha}{2}}\right\}=1-\alpha$$

故

$$\left(\overline{X}-\frac{\sigma}{\sqrt{n}}u_{\frac{\alpha}{2}},\overline{X}+\frac{\sigma}{\sqrt{n}}u_{\frac{\alpha}{2}}\right) \tag{4-2}$$

为μ的置信水平$1-\alpha$的置信区间。

【例 4-18】 一车间生产的滚珠直径服从正态分布，从某天的产品里随机抽取 6 个，测得直径为（单位：mm）14.6，15.1，14.9，14.8，15.2，15.1。若该天产品直径的方差$\sigma^2=0.06$，求该天生产的滚珠的平均直径μ的置信区间（$\alpha=0.01$；$\alpha=0.05$）。

解：因为$\sigma^2=0.06$，由式（4-2）知μ的$1-\alpha$的置信区间为$\left(\overline{X}-\frac{\sigma}{\sqrt{n}}u_{\frac{\alpha}{2}},\overline{X}+\frac{\sigma}{\sqrt{n}}u_{\frac{\alpha}{2}}\right)$。

当$\alpha=0.01$时，查正态分布表得$u_{\alpha/2}=2.58$，计算得$\overline{x}=14.95$，将$\overline{x}=14.95$，$\sigma^2=0.06$，$n=6$，$u_{0.005}=2.58$代入上述置信区间，得μ的99%的置信区间为

$$\left(14.95-2.58\sqrt{\frac{0.06}{6}},14.95+2.58\sqrt{\frac{0.06}{6}}\right)=(14.69,15.21)$$

当$\alpha=0.05$时，查正态分布表得$u_{\alpha/2}=1.96$，求得μ的95%的置信区间为（14.75, 15.15）。

（2）σ^2未知时，均值μ的置信区间

这时，自然地会想到以样本标准差S代替总体均方差σ，由正态公式知选取统计量

$$T=\frac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}}\sim t(n-1)$$

对给定的数α，由

$$P\left\{|T|<t_{\frac{\alpha}{2}}(n-1)\right\}=1-\alpha$$

查 t 分布表，得$t_{\frac{\alpha}{2}}(n-1)$，解不等式得$\overline{X}-\frac{S}{\sqrt{n}}t_{\frac{\alpha}{2}}(n-1)<\mu<\overline{X}+\frac{S}{\sqrt{n}}t_{\frac{\alpha}{2}}(n-1)$，即$\mu$的置信水平$1-\alpha$的置信区间为

$$\left(\overline{X}-\frac{S}{\sqrt{n}}t_{\frac{\alpha}{2}}(n-1),\overline{X}+\frac{S}{\sqrt{n}}t_{\frac{\alpha}{2}}(n-1)\right)$$

简记为

$$\bar{X} \pm \frac{S}{\sqrt{n}} t_{\frac{\alpha}{2}}(n-1) \tag{4-3}$$

在实际问题中，很难找到总体均值未知，但方差已知的情况。通常情况下，均值和方差都要通过样本进行估计，故式（4-3）比式（4-2）更实用。

【例 4-19】 水体中的污水和工业污染会通过减少水中被溶解的氧气而影响水体的水质，生物的生长与生存依赖于氧气。两个月内，从污水处理厂下游 1mile（1mile=1609.344m）处的一条小河里取 8 份水样。检测水样里溶解的氧气含量，数据见表 4-6。

表 4-6　水样中的氧气含量

水样/份	1	2	3	4	5	6	7	8
氧气含量 × 10^{-6}	5.1	4.9	5.6	4.2	4.8	4.5	5.3	5.2

根据最近的研究，为了保证鱼的生存，水中溶解的氧气的平均体积含量需达到 5×10^{-6}，即试求两个月期间平均氧气含量的 95%的置信区间（假定样本来自正态总体）。

解： σ^2 未知，所以由式（4-3）知 μ 的置信水平 $1-\alpha$ 的置信区间为 $\left(\bar{X}-\frac{S}{\sqrt{n}}t_{\frac{\alpha}{2}}(n-1),\right.$ $\left.\bar{X}+\frac{S}{\sqrt{n}}t_{\frac{\alpha}{2}}(n-1)\right)$，

由已知 $n=8$，$1-\alpha=0.95$，查附录 C 得 $t_{0.025}(7)=2.365$，由样本计算得 $\bar{x}=4.95$，$S=0.45$，故 μ 的置信水平 $1-\alpha$ 的置信区间为（4.78, 5.12）。

2．单正态总体方差的区间估计

设总体 $X\sim N(\mu,\sigma^2)$，$X_1,X_2,\cdots,X_n$ 是 X 的样本，求 σ^2 的置信水平 $1-\alpha$ 的置信区间。由 χ^2 分布式知选取统计量

$$\frac{(n-1)S^2}{\sigma^2}\sim\chi^2(n-1)$$

对给定的 α，取 χ^2 分布分位点 $\chi^2_{\frac{\alpha}{2}}(n)$ 和 $\chi^2_{1-\frac{\alpha}{2}}(n)$，使

$$P\left\{\chi^2_{1-\frac{\alpha}{2}}(n-1)<\frac{(n-1)S^2}{\sigma^2}<\chi^2_{\frac{\alpha}{2}}(n-1)\right\}=1-\alpha$$

从而得到 σ^2 的置信水平 $1-\alpha$ 的置信区间为

$$\left(\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}(n-1)},\frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}\right)$$

【例 4-20】 从自动机床加工的同类零件中随机地抽取 10 件，测得其长度为（单位：mm）12.15，12.12，12.10，12.28，12.09，12.16，12.03，12.01，12.06，12.11。假定样本来自正态总体，试求方差 σ^2 的 95%的置信区间。

解： 已知 $\alpha=0.05$，查附录 B 得

$$\chi^2_{1-\frac{\alpha}{2}}(n-1)=\chi^2_{0.975}(9)=2.7，\quad \chi^2_{\frac{\alpha}{2}}(n-1)=\chi^2_{0.025}(9)=19.023$$

又由已知数据算得$S=0.076$，于是

$$\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}=\frac{9\times0.076^2}{19.023}=0.003,\quad \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}=\frac{9\times0.076^2}{2.7}=0.019$$



所以，方差σ^2的95%的置信区间是$\left(\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}(n-1)},\frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}\right)$=（0.003, 0.019）。

4.4.3 单侧置信区间

在上面的内容中，对于未知参数θ，给出两个统计量$\underline{\theta}$和$\overline{\theta}$，得到θ的置信区间为$(\underline{\theta},\overline{\theta})$的形式。在某些实际应用中，常常只关心参数的上限或下限。例如，对于设备、元件的寿命来说，只关心平均寿命θ至少是多少（θ的下限）。与之相反，在考虑化学药品中杂质含量时，关心的却是平均杂质含量θ'最多是多少（θ'的上限）。这就引出了单侧置信区间的概念。

对于给定值$\alpha(0<\alpha<1)$，若由样本$X_1,X_2,\cdots,X_n$确定的统计量$\underline{\theta}(X_1,X_2,\cdots,X_n)$满足对任意$\theta$有

$$P\{\theta>\underline{\theta}\}=1-\alpha$$

则称随机区间$(\underline{\theta}(X_1,X_2,\cdots,X_n),+\infty)$是$\theta$的置信水平为$1-\alpha$的下侧置信区间，称$\underline{\theta}(X_1,X_2,\cdots,X_n)$是置信水平为$1-\alpha$的单侧置信下限。

若统计量$\overline{\theta}(X_1,X_2,\cdots,X_n)$满足对任意$\theta$有

$$P\{\theta<\overline{\theta}\}=1-\alpha$$

则称随机区间$(-\infty,\overline{\theta}(X_1,X_2,\cdots,X_n))$是$\theta$的置信水平为$1-\alpha$的上侧置信区间，称$\overline{\theta}(X_1,X_2,\cdots,X_n)$是置信水平为$1-\alpha$的单侧置信上限。

【例 4-21】 已知某地区农户人均生产蔬菜量$X\sim N(\mu,\sigma^2)$，现随机抽取 9 户，得人均生产的蔬菜量（单位：kg）为 75，143，156，340，400，287，256，244，249。问该地区农户人均生产蔬菜最多为多少（$\alpha=0.05$）？

解：这里总体方差未知，求均值的置信上限。选取统计量

$$T=\frac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}}\sim t(n-1)$$

对给定的α，有

$$P\left\{\left|\frac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}}\right|>t_\alpha(n-1)\right\}=1-\alpha$$

则考虑单侧置信上限时有

$$P\left\{\mu<\overline{X}+\frac{S}{\sqrt{n}}t_\alpha(n-1)\right\}=1-\alpha$$

容易得到 μ 的置信上限为 $\overline{\mu}=\overline{X}+\frac{S}{\sqrt{n}}t_\alpha(n-1)$。

由样本算得 $\overline{X}=239\text{kg}$，$S=101\text{kg}$，查 t 分布表得 $t_\alpha(n-1)=t_{0.05}(8)=1.86$，于是 μ 的95%的置信上限为 $\overline{\mu}=\overline{X}+\frac{S}{\sqrt{n}}t_\alpha(n-1)=239+\frac{101}{\sqrt{9}}\times1.86=302$。

结果表明，该地区农户人均生产蔬菜最多是 302kg，这一估计的置信水平为 95%。

4.5 概率分布的统计特征

4.5.1 概率密度和累积分布密度

对每一种分布，统计工具箱提供了计算给定变量 x 的概率值的函数，其形式为 xxxpdf。下面通过示例来展示其用法。

【例 4-22】 计算 $x=50$ 二项式分布的概率。

其实现的 MATLAB 程序代码如下：

```
>> clear all;
%设置二项式分布的参数
N=100;p=0.5;
%x 的值
x=50;
%计算概率
y=binopdf(x,N,p)
```

运行程序，输出如下：

```
y =      0.0796
```

假设 f 是随机变量 X 的概率密度函数，则其相应的累积分布函数定义为：

$$F(X)=P\{X\leqslant x\}=\int_{-\infty}^{x}f(t)\mathrm{d}t$$

统计工具箱中提供了计算不同累积分布的函数，其形式为 xxcdf。下面举例来说明这类函数的用法。

expcdf 函数的调用格式如下：

```
P=expcdf(X, MU)
```

expcdf 函数用于计算指数累积分布。其中，MU 是指数分布的参数；P 为返回的概率累积分布。

其相关函数有：cdf, expfit, expinv, exppdf, exprnd, expstat。

【例 4-23】 计算指数分布随机变量小于均值的概率。

其实现的 MATLAB 程序代码如下：

```
>> clear all;
```

```
%指数分布的参数
mu=10;
%变量
x=mu;
%计算累积概率
p=expcdf(x,mu)
```



运行程序，输出如下：

```
p =     0.6321
```

4.5.2 概率分布的均值和方差

通过前面的学习，知道在数理统计中，随机变量常见的数字特征包括均值和方差，而统计工具箱中提供了计算不同分布均值和方差的函数，其形式为 xxxstat，而且还提供了显示均值及其置信区间位置的 grpstats 函数。下面进行举例说明。

（1）wblstat 函数

其调用格式如下：

```
[M,V]=wblstat(A,B)
```

wblstat 函数用于计算 Weibull 分布的均值和方差。其中，A，B 是 Weibull 分布的参数；M 为返回均值；V 为返回方差。

其相关函数有：wblcdf, wblfit, wblinv, wbllike, wblpdf, wblplot, wblrnd。

（2）grpstats 函数

其调用格式如下：

```
means=grpstats(X, group)
[means, sem, counts, name]=grpstats(X, groud)
grpstats(x,group,alpha)
```

grpstats 函数用于计算每组的统计量。其中，X 是分析的矩阵，每一列是一组数据；group 是组的索引；*alpha* 是置信水平；*means* 返回每一列的均值；无输出参数时，显示每个均值 $100(1-\alpha)$的置信区间。

【例 4-24】 计算 Weibull 分布的均值和方差。

其实现的 MATLAB 程序代码如下：

```
>> clear all;
%设置 Weibull 分布的参数
a=0.4;b=3;
%计算均值和方差
[M,V]=wblstat(a,b)
%产生两组数据
len=30;
group=unidrnd(2,len,1);
true_mean=1:2;
```

```
true_mean=true_mean(ones(len,1),:);
y=wblrnd(true_mean,1);
%显示置信水平
alpha0=0.05;
means=grpstats(y,group,alpha0)
```

运行程序，输出如下：

```
M =      0.3572
V =      0.0169
means =
    0.6358      1.1390
    1.2771      1.2482
```

两组数据的均值及置信区间位置如图 4-16 所示。

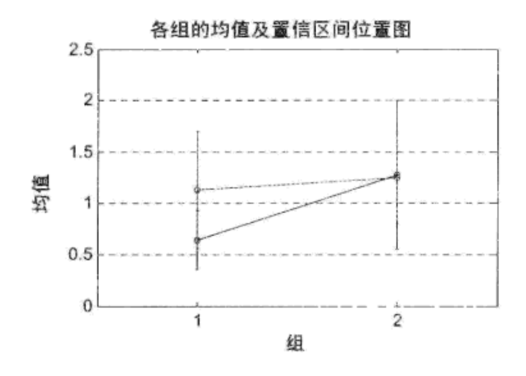


图 4-16　均值及置信区间位置图





第 5 章 统计检验方法——假设检验

统计推断的另一类重要问题是假设检验问题。先对总体的某个未知参数或总体的分布形式作某种假设，然后由所抽取的样本提供的信息，构造合适的统计量，对所提出的假设进行检验，以作出统计判断。是接受假设还是拒绝假设，这类统计推断问题称为假设检验问题。

5.1 假设检验概述

在统计应用中会遇到如下类型的问题。

【例 5-1】 一台自动车床在正常工作的情况下加工出的零件直径服从正态分布，零件规格是：标准直径为 5cm，允许的最大加工误差为 0.2cm。某日开工后，技术人员进行例行检查，以判断该车床工作是否正常。

这是一个生产设备运行稳定性的监督问题。在工业生产中监督设备的运行稳定性，通常的做法如下：

① 进行例行监督检查。此时，往往假定设备的工作是正常的，然后每隔一段时间随机抽查几个产品的控制指标（如零件直径），如果没有发现异常情况，就认为生产是正常的。如果发现产品的质量有大的变动，超过了允许的限度，则认为生产不正常而需要停机检修。用统计语言描述就是，假设变量的分布形态已知，判断关于分布参数的一些已知信息是否为真，即进行变量分布参数的假设检验。

② 在生产环境发生变化，如设备大修或工艺改变等情况下，需要判断设备的运行是否符合正常状态要求，这不仅涉及①中所述的参数检验问题，而且首先要判断产品的控制指标的概率分布是否与要求的一样。用统计语言描述就是，对变量的分布形态已有先验的知识，如变量曾经或者应该服从正态分布、威布尔分布等，判断目前的情况是否如此。

假设检验是一类重要的、应用广泛的统计推断技术。本节讲解假设检验的基本思想、方法和步骤等问题。

5.1.1 假设检验的逻辑

仍以例 5-1 中的问题为例，讲解假设检验的基本思想和方法。假设这台自动车床的工作是正常的，零件直径服从正态分布，进行例行的质量检查。假定从一天的产品中抽查 50 个，分别测量直径，算得 $\overline{X}=4.8\text{cm}$。据此来推断这台自动车床当天的生产是否正常。

这就是变量分布参数的假设检验问题。

在假设检验问题的分析与推理中，首先要明确待检验的命题 H_0，称为统计假设（也叫原假设或零假设，称与之对立的假设 H_1 为备择假设），然后由抽样结果来检查这个假设是否可信、是否能够成立，从而做出拒绝还是接受这个假设的决策。

在例 5-1 中，一天中生产的零件的平均直径是一个随机变量 X，已知 X 服从正态分布。

现在想知道，这一天生产的所有零件的直径 $E(X)=\mu$ 是否符合标准要求，即 $\mu=5$ 是否成立。如果 $\mu=5$，说明生产正常；否则，说明生产不正常。

于是，设原假设 H_0：$\mu=5$；备择假设 H_1：$\mu\neq 5$。

怎样来判定 H_0 是否为真呢？由于 $X\sim N(\mu,\sigma^2)$，即 μ 是零件直径的期望值，而样本均值 $\bar{X}$ 是 μ 的性能优良的估计量，H_0 是否为真的判断可以通过定量分析二者的信息差异得到。现在 $\bar{X}=4.8$，而要求 $\mu=5$，其间存在差异 $\bar{X}-\mu=-0.2$，于是 H_0 是否为真取决于这个差异的性质。

① 差异可能是由随机因素引起的，称为抽样误差或随机误差，这种误差反映偶然的、非本质的因素引起的随机波动。

② 差异不是由随机因素引起的，它反映事物的本质差别（反映这天生产的零件的平均直径同标准直径不同），称为系统误差。

那么，这个抽样结果究竟是偶然性在起作用，还是该天生产不正常所造成的？这就需要给出一个量的界限，即给出一个小的正数 δ。如果 $\left|\bar{X}-\mu\right|<\delta$，则认为是随机性的差异，或者用统计学上的术语称差异不够显著；如果 $\left|\bar{X}-\mu\right|\geqslant\delta$，则认为不是随机性的差异，或者说差异显著。

于是，问题转化为如何确定这个正数 δ。容易想到，可以采用区间估计中的大概率置信准则

$$P\left\{\left|\bar{X}-\mu\right|<\delta\right\}\geqslant 1-\alpha$$

来确定这个量的界限 δ。

但是，这里产生了一个问题：$\bar{X}$ 是一个随机变量，用 $\bar{X}$ 的观测值说明命题 H_0：$\mu=5$ 的真假是一种事实验证，若在一次抽样中 $\left|\bar{X}-\mu\right|<\delta$，只能增加人们对命题 H_0 的信心，即使是 100 次的验证都支持命题 H_0，但是仍不能令人相信命题 H_0 是真的。

如果注意到当 $X\sim N(\mu,\sigma^2)$ 时，有 $X\sim N\left(\mu,\dfrac{\sigma^2}{n}\right)$，即当 H_0 为真时，$\bar{X}$ 的观测值不应过于偏离 $\mu=5$，即事件 $\left\{\left|\bar{X}-\mu\right|\geqslant\delta\right\}$ 应当是一个小概率事件，不妨记为

$$P\left\{\left|\bar{X}-\mu\right|\geqslant\delta\right\}\leqslant\alpha$$

称为检验准则，其中 α 是一个很小的正数，称为显著性水平。小概率事件在一次试验中基本上不会发生。如果在一次抽样中，$\bar{X}$ 的样本观测值 $\bar{x}\in W$，即 $\bar{X}$ 的观测值过于偏离 $\mu=5$，试验结果与前提假设不相符，则使人们不能不怀疑作为这个小概率事件前提的命题 H_0 的正确性。这里的集合 W 称为 H_0 的拒绝域。如果一个概率很小的事件在一次试验中发生了，则人们认为命题 H_0 不真的理由比承认命题 H_0 为真更为充分。也就是说，在假设检验问题中，采用伺机否定 H_0 的思维逻辑比执意支持 H_0 的思维逻辑更有说服力。

把伺机否定 H_0 的思维过程中使用的推理方法称为概率反证法，它不同于一般的反证法。一般的反证法如果在原假设下导出的结论自相矛盾或与事实矛盾，则完全绝对地推翻原假设；而概率反证法的结论不是绝对的，只是认为结论正确的把握较大，不排除犯错误的可能。

假设检验推理方法是概率反证法，其推理逻辑是：如果原假设 H_0 是对的，而能够验证 H_0 为真的某个统计量落入某个约定的区域 W 是小概率事件，而小概率事件在一次试验中基本上不会发生。如果该统计量的一次实测值落入区域 W。也就是说，原假设成立下的小概率事件在一次试验中发生了，那么就以较充分的理论认为原假设不可信而否定它，否则就不能否定原假设（只好接受它）。不否定原假设并不是原假设一定对，而只是说差异还不够显著，还没有达到足以否定原假设的程度。



5.1.2 假设检验的步骤

假设检验的基本步骤如下。

第一步，提出原假设 H_0 及备择假设 H_1。

原假设是对问题的标准统计描述，是待验证的命题。备择假设则是原假设的对立命题，是否定原假设结论时的统计描述。

例 5-1 中，原假设 H_0： $\mu=\mu_0=5$；备择假设 H_1： $\mu\neq\mu_0$。

称这类假设检验为双侧假设检验，有时还会提出下述形式的假设：

$$H_0: \ \mu\leqslant\mu_0; H_1: \ \mu>\mu$$

或

$$H_0: \ \mu\geqslant\mu_0; H_1: \ \mu<\mu_0$$

称这类假设检验为单侧假设检验。

此外要注意，对于一个实际问题，原假设通常有两种提法，即原假设和备择假设可以互换。应该如何提取原假设呢？这里给出一个原则性的建议：在实际问题中，往往把系统早已存在或样本信息明显支持的状态、不宜轻易否定的命题作为原假设 H_0，或者说把希望得到或反映系统新变化的结论作为备择假设 H_1。

第二步，选取一个适当的检验统计量 T，并写出相应的检验准则。

如例 5-1 中，检验统计量为 $\bar{X}$，检验准则是 $P\left\{\left|\bar{X}-0.5\right|\geqslant\delta\right\}\leqslant\alpha$。

在这一环节应当注意，在 H_0 成立的条件下，所选定的检验统计量 T 的概率分布（或近似分布）应当是已知的，如例 5-1 中，若 H_0 成立，即 $X\sim N(0.5,0.2^2)$ 时，有 $\bar{X}\sim N(5,0.0008)$。

拒绝域的临界值的计算依赖于检验统计量的概率分布。有时为了便于计算，特别是查表计算的情况下，需要对检验统计量进行分布形态规范化、标准化或渐近正态化变换。如例 5-1 中，通常需要将检验统计量 $\bar{X}$ 变换为 $U=\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}$，在 H_0： $\mu=5$ 成立时，$U\sim N(0,1)$。

第三步，给定显著性水平 α，并求出 H_0 的拒绝域 W。

如例 5-1 中，给定的显著性水平 $\alpha=0.05$，由检验准则

$$P\left\{\left|\bar{X}-0.5\right|\geqslant\delta\right\}\leqslant\alpha$$

可得

$$P\{\bar{X}\leqslant 0.5-\delta\}+P\{\bar{X}\geqslant 0.5+\delta\}\leqslant 0.05$$

即

$$W=(-\infty,a]\cup[b,+\infty)$$

其中，$a=0.5-\delta$，$b=0.5+\delta$。通常用等分配置显著性水平的方法确定拒绝域的临界值，即

$$P\{\bar{X}\leqslant 0.5-\delta\}\leqslant 0.025，\quad P\{\bar{X}\geqslant 0.5+\delta\}\leqslant 0.025$$

进而，根据 $\bar{X}\sim N(5,0.0008)$，计算拒绝域的临界值。

其实现的 MATLAB 程序代码如下：

```
>> a=norminv(0.025,5,0.0008)
b=norminv(0.975,5,0.0008)
```

运行程序，输出如下：

```
a =      4.9984
b =      5.0016
```

即原假设 H_0 的拒绝域为 $W=(-\infty,4.9984]\cup[5.0016,+\infty)$。

第四步，由样本计算出检验统计量 T 的实测值，判断其是否落入拒绝域。

若实测值落入拒绝域，则认为差异显著而否定原假设 H_0；否则，认为差异不显著而不能否定原假设，即保留（接受）原假设 H_0。

如例 5-1 中，$\bar{X}=4.8\in W$，故否定原假设 H_0，即认为这天生产不正常，需检修。

上面作出的否定原假设的判断，判断正确的置信水平为 0.95，判断错误的风险概率为 0.05。

5.1.3 检验的 p 值

在假设检验问题中，得出结论的依据是检验统计量 T 的观测值 t 是否落入原假设 H_0 的拒绝域 W。如果 $t\in W$，则拒绝原假设 H_0，否则保留原假设 H_0。这种非此即彼的结论有一个缺点，即结论不能反映由当前的样本信息拒绝（或保留）原假设的理由是否充分。具体地讲，检验统计量 T 的观测值 t 虽然落入拒绝域 W，但其距离 W 的临界值有多远？如例 5-1 中，W 的左侧临界值为 4.9984，检验统计量 $\bar{X}$ 的值为 4.8，小于 4.998，落入 W，故拒绝原假设 H_0。问题是：依据 4.8<4.998 得出结论，理由是否勉强？对此最好有一个数量上的刻画。“检验的 p 值”能够满足人们的这种要求。

定义 5-1（检验的 p 值） 设原假设为 H_0，T 是检验统计量，其观测值为 t，H_0 的拒绝域为 W，则称如下定义的概率 P 为原假设 H_0 的检验的 p 值。

若 $W=\{T\mid T\geqslant c\}$，则 $p=P\{T\geqslant t\mid H_0为真\}$。

若 $W=\{T\mid T\leqslant c\}$，则 $p=P\{T\leqslant t\mid H_0为真\}$。

若 $W=\{T\mid T\leqslant c_1或T\geqslant c_2\}$，则

① 当 t 值较小（偏左取值）时，$p=2P\{T\leqslant t\mid H_0为真\}$。

② 当 t 值较大（偏右取值）时，$p=2P\{T\geqslant t\mid H_0为真\}$。

在统计实践中，人们并不事先指定显著性水平 α 的值，而是很方便地利用上面定义的 p

值。对于任意大于 p 值的显著性水平，人们可以拒绝原假设，但不能在任何小于它的显著性水平下拒绝原假设。p 值是利用样本数据能够作出拒绝原假设的最小的显著性水平。

【例 5-2】 某人有 4 枚不同的硬币，他怀疑这 4 枚硬币的均匀性不同，想通过抛掷硬币观察出现正面的次数来鉴别硬币的均匀性。于是进行了掷币试验，4 枚硬币各抛掷 100 次，并记录了出现正面的次数，结果见表 5-1。



表 5-1 硬币正面次数表

硬币编号	1	2	3	4
出现正面的次数	50	55	60	65

分析：设在 100 次抛掷中每枚硬币出现正面的次数为 X_i，每次抛掷出现反面的概率分别为 $p_i(i=1,2,3,4)$，则 $X_i \sim b(100,p_i)$。检验的原假设为

$$H_0^{(i)}:\ p_i = 0.5\ （硬币是均匀的），\ i=1,2,3,4$$

在 H_0 为真的假设下，即 $X_i \sim b(100,0.5)$，出现正面的平均次数为 $E(X_i)=100\times 0.5=50$。由于实测出现正面的次数均不小于 50，故可作单侧检验，即备择假设为

$$H_1^{(i)}:\ p_i > p_0 = 0.5,\ \ i=1,2,3,4$$

在显著性水平 α 下，检验准则是

$$P\{X_i - 50 \geqslant \delta\} \leqslant \alpha$$

下面，利用 MATLAB 分别来求 H_0 的拒绝域和检验的 p 值。

① 求拒绝域，这里指定显著性水平 $\alpha = 0.05$。由于检验统计量服从相同的分布，故对每种硬币而言，原假设的拒绝域是相同的。

其实现的 MATLAB 程序代码如下：

```
>> clear;
Wlower=binoinv(0.95,100,0.5)   %求拒绝域的临界值 50+ δ
```

运行程序，输出如下：

```
Wlower =     58
```

② 求对每枚硬币进行检验的 p 值：$p_i = P\{X_i > x_i\}\ (i=1,2,3,4)$。

其实现的 MATLAB 程序代码如下：

```
>> clear;
p1=1-binocdf(50,100,0.5);
p2=1-binocdf(55,100,0.5);
p3=1-binocdf(60,100,0.5);
p4=1-binocdf(65,100,0.5);
p=[p1,p2,p3,p4]
```

运行程序，输出如下：

```
p =
```

0.4602　　0.1356　　0.0176　　0.0009

根据上述计算可知，在 0.05 显著性水平下，检验认为第一和第二两种硬币是均匀的，而第三和第四两种硬币不是均匀的。

如果改变显著性水平，则需要重新计算拒绝域的临界值，但是利用检验的 p 值进行决策则不必重新计算，应用起来更为灵活方便。在 0.05 显著性水平下，检验的 p 值表明不必质疑第一种硬币均匀而第四种硬币不均匀的结论。如果严格均匀性的标准，即增大显著性水平（更容易拒绝原假设），如取 0.15，则统计推断不能认为第二种硬币是均匀的；如果放宽均匀性的标准，即减小显著性水平（不容易拒绝原假设），如取 0.01，则统计推断认为第三种硬币是均匀的。

5.1.4 假设检验错误与势函数

在假设检验方法的应用中，必须注重检验的结果是否与实际情况吻合。换句话说，假设检验是可能犯错的。在作出否定原假设的判断时，可能犯如下两类错误。

① 第一类错误。H_0 本来是正确的，但由于随机性使检验统计量的观测值落入拒绝域（小概率事件并非不可能发生），依检验规则应当否定原假设。这时的结论犯了“以真为假”的错误，即否定了正确的原假设。

显然，5.1.1 节中讲解的检验准则是对检验中犯第一类错误的概率控制，即

$$P\{否定H_0 \mid H_0为真\} = P\{第一类错误\} = \alpha$$

α 为事先给定的显著性水平。

② 第二类错误。如果原假设 H_0 是错误的，同样由于随机性使检验统计量的观测值没有落入拒绝域，依检验规则不能否定原假设。这时的结论犯了“以假为真”的错误，即接受了错误的原假设。犯第二类错误的概率记为

$$P\{不否定H_0 \mid H_0为假\} = P\{第二类错误\} = \beta$$

或

$$P\{接受H_0 \mid H_1为真\} = P\{第二类错误\} = \beta$$

我们希望检验的结论使犯两类错误的概率同时都很小，最好是全为 0，但这是一个两难问题，当样本容量给定后，犯这两类错误的概率就不能同时被控制了。为了说明这种两难性，引入检验的势函数的概念。

定义 5-2（检验的势函数）　设 Θ 为 θ 的参数空间，$\Theta_0 \cup \Theta_1 = \Theta$ 且 $\Theta_0 \cap \Theta_1 = \varnothing$。检验的原假设 H_0：$\theta \in \Theta_0$（备择假设为 H_1：$\theta \in \Theta_1$）的拒绝域为 W，则检验统计量 T 的观测值落入拒绝域 W 的概率

$$g(\theta) = P\{T \in W\},\quad \theta \in \Theta$$

称为该检验的势函数。

实质上，势函数是对犯第一类错误的概率 α ($\alpha(\theta)$) 和犯第二类错误的概率 β ($\beta(\theta)$) 的统一描述，是参数 θ 的函数。其关系式为

$$g(\theta) = \begin{cases} \alpha(\theta), & \theta \in \Theta_0 \\ 1-\beta(\theta), & \theta \in \Theta_1 \end{cases}$$



或

$$\alpha(\theta)=g(\theta),\quad \theta\in\Theta_0$$
$$\beta(\theta)=1-g(\theta),\quad \theta\in\Theta_1$$

为表述简单，在变量 $X\sim N(\mu,\sigma^2)$，σ^2 已知的条件下，以检验

$$H_0:\ \mu\geqslant\mu_0;\quad H_1:\ \mu<\mu_0$$

为例，对这一结论进行说明。同例 5-1，这里 H_0 的检验统计量仍为 $\bar{X}$，拒绝域 $W=(-\infty,c]$，于是

$$g(\mu)=P\{\bar{X}\leqslant c\}=P\left\{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\leqslant\frac{c-\mu}{\sigma/\sqrt{n}}\right\}=\Phi\left(\frac{c-\mu}{\sigma/\sqrt{n}}\right)$$

又由犯第一（二）类错误的概率 $\alpha(\beta)$ 的定义可知：

当 $\mu\geqslant\mu_0$ 时，$g(\mu)=P\{\bar{X}\in W\}=P\{否定H_0\mid H_0为真\}=\alpha$，即 α 是 μ 的函数。

当 $\mu<\mu_0$ 时，$g(\mu)=P\{\bar{X}\in W\}=P\{否定H_0\mid H_1为真\}=1-P(接受H_0\mid H_1为真)=1-\beta$，即 β 也是 μ 的函数。

显然，犯两类错误的概率可统一由势函数表示，即

$$\alpha(\mu)=g(\mu)=\Phi\left(\frac{c-\mu}{\sigma/\sqrt{n}}\right),\quad \mu\geqslant\mu_0$$

$$\beta(\mu)=1-g(\mu)=1-\Phi\left(\frac{c-\mu}{\sigma/\sqrt{n}}\right),\quad \mu<\mu_0$$

由这两个式子可以看出（σ 和 n 是确定的，$\Phi\left(\frac{c-\mu}{\sigma/\sqrt{n}}\right)$ 是 c 的单调增函数），要使 α 减小，应使 $\Phi\left(\frac{c-\mu}{\sigma/\sqrt{n}}\right)$ 中的 c 变小，此时导致 $1-\Phi\left(\frac{c-\mu}{\sigma/\sqrt{n}}\right)$ 变大，即 β 变大；反之，要使 β 减小，应使 $1-\Phi\left(\frac{c-\mu}{\sigma/\sqrt{n}}\right)$ 变小，此时导致 c 变大，即 α 变大。这说明在假设检验的过程中，在给定样本容量的条件下，人们不可能使犯两类错误的概率同时都很小，即 α 与 β 之间一个变小必然导致另一个变大。

因此，在假设检验的实际应用中，通常人们只能控制犯第一类错误的概率，即根据实际情况，通过控制显著性水平 α 的大小来减少犯错误的可能性。这种做法通常称为显著性检验。

在显著性检验过程中，当人们宁可“以假为真”而不愿“以真为假”时，则应把 α 取得很小，如 $\alpha=0.01$。反之，则应把 α 取得大些，如 $\alpha=0.1$，折中的取法是 $\alpha=0.05$。例如，某药品含有毒性，必须严格控制不得超过规定的指标。如果设原假设为产品不合格（毒性超过某一标准），则应把 α 取得很小，这样才能保证用药的安全，当然难免会把一些合格品当成废品处理了。在另一些情况下正好相反，如检查袋装食品的质量，就没有必要那么严格，如果原假设为产品不合格（质量低于某标准），可以把 α 取得稍大些。不管在什么情况下，为了保证 β 不致于太大，样本容量都不应太小。

5.1.5 假设检验与区间估计的关系

假设检验与区间估计是两种最重要的统计推断形式，这两者初看起来好像完全不同，其实两者之间有一定的联系。利用区间估计可建立假设检验，反之亦然。下面仍通过例 5-1 作简要说明。

设总体 $X \sim N(\mu,\sigma^2)$，σ^2 已知，若求 μ 的区间估计，应选择统计量

$$U=\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

按置信水平 $1-\alpha$ 确定一个大概率事件

$$P\left\{\left|\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right|<u_{1-\alpha/2}\right\}=1-\alpha$$

由此，得到 μ 的置信水平为 $1-\alpha$ 的区间估计为

$$\left(\bar{X}-u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}},\bar{X}+u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

这个区间估计恰好是原假设 H_0： $\mu=\mu_0$ 的一个接受区域，显著性水平为 α。

问题是：如果检验假设

$$H_0:\ \mu=\mu_0;\ H_1:\ \mu\neq\mu_0$$

选取的统计量是

$$U=\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$$

对给定的显著性水平 α，得到小概率事件

$$P\left\{\left|\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}\right|\geqslant u_{1-\alpha/2}\right\}=\alpha$$

由实测值 $\left|\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}\right|\geqslant u_{1-\alpha/2}$ 是否成立，决定是否拒绝原假设。

拒绝域为 $\left|\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}\right|\geqslant u_{1-\alpha/2}$，则接受域为 $\left|\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}\right|<u_{1-\alpha/2}$，再把 μ_0 改为 μ，那么结果正是 μ 的区间估计，置信水平为 $1-\alpha$。

需要注意的是，假设检验和区间估计的结果在解释上是有差别的。

例如，在检验 H_0： $\mu=\mu_0=0$（显著性水平为 α）的同时对 μ 作区间估计（置信水平为 $1-\alpha$），可能会出现以下几种情况。

① 检验的结论与区间估计一致。如检验接受 H_0，区间估计为（–0.001, 0.001）。按假设检验，应接受 $\mu=0$；按区间估计，μ 可能取到的最大值和最小值都很接近 0，两者的解释一致。

② 区间估计强化了检验的结论。如果检验拒绝 H_0，区间估计为（1000, 2000）。按假设检验，应拒绝 $\mu=0$；按区间估计，区间中不包含 0，即 0 不看做 μ 的一个可能值，而且区

间的最小值也有 1000，与 0 相去甚远，故认为$\mu \neq 0$的理由很充分，区间估计的结论加强了假设检验的结论。

③ 检验的结论与区间估计不协调。如检验拒绝H_0，区间估计为（0.001, 0.002)。按假设检验，应拒绝$\mu=0$；按区间估计，区间中不包含 0，从这个方面看两者一致。可是细看这个区间，就发现整个区间在 0 的附近，因此实质上可以认为μ就是 0。这样，区间估计的结论（在实质上）就与假设检验不同了。又如检验接受H_0，区间估计为（–1000, 1500)。按假设检验，应接受$\mu=0$；按区间估计，区间中包含 0，即 0 是μ的一个可能值，在这一点上与假设检验的结论一致。可是细看这个区间，最大可以到 1500，最小可以到–1000，这中间哪一个值都有可能。因此，从区间估计的角度来看，实在没有多大把握认为μ的取值都在 0 附近，这就与假设检验的结论不大协调了。



由此例可以看出，统计上的结论一定要注意其实质含义，如只停留在表面，就有可能被引入误区。

5.2 单正态总体的假设检验

正态总体$N(\mu,\sigma^2)$的假设检验问题主要有下列几种：

- 已知方差σ^2，检验零假设H_0： $\mu=\mu_0$（μ_0为已知数）。
- 未知方差σ^2，检验零假设H_0： $\mu=\mu_0$（μ_0为已知数）。
- 未知期望μ（均值)，检验零假设H_0： $\sigma^2=\sigma_0^2$（σ_0为已知数）。
- 未知期望μ（均值)，检验零假设H_0： $\sigma^2 \leqslant \sigma_0^2$（$\sigma_0$为已知数）。

5.2.1 总体均值的检验

1.已知方差σ^2，关于μ的检验（u检验法）

已知方差σ^2，关于μ的检验主要有双边检验和单边检验两种。

- 双边检验：H_0： $\mu=\mu_0$； H_1： $\mu\neq\mu_0$。
- 单边检验：H_0： $\mu=\mu_0$； H_1： $\mu>\mu_0$（或H_1： $\mu<\mu_0$）。

现就两种检验方法介绍如下。

设$X_1,X_2,\cdots,X_N$是正态总体$\overline{X}\sim N(\mu,\sigma^2)$的一个样本，其中$\mu$未知，$\sigma^2=\sigma_0^2$（已知）。用样本检验假设。

$$H_0：\mu=\mu_0\ （\mu_0\text{为已知数}）；\ H_1：\mu\neq\mu_0$$

当H_0成立时，检验统计量U满足

$$U=\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}\sim N(0,1)$$

对给定显著性水平α，查标准正态分布表，得临界值$z_{\frac{\alpha}{2}}$，使得

$$\Phi\left(z_{\frac{\alpha}{2}}\right)=1-\frac{\alpha}{2}$$

可以算出 $P\left\{|U|\geqslant z_{\frac{\alpha}{2}}\right\}=\alpha$，由此得 H_0 的拒绝域 $\left(-\infty,-z_{\frac{\alpha}{2}}\right]\cup\left[z_{\frac{\alpha}{2}},+\infty\right)$ 和 H_0 的相容域 $\left(-z_{\frac{\alpha}{2}},z_{\frac{\alpha}{2}}\right)$。由样本值 $x_1,x_2,\cdots,x_n$ 计算检验统计量 U_0 的值。

若 $|U|\geqslant z_{\frac{\alpha}{2}}$（即落在拒绝域中），则拒绝 H_0，接受 H_1；若 $|U|<z_{\frac{\alpha}{2}}$（即落在相容域中），则接受 H_0。

这种检验统计量服从 $N(0,1)$。把通过检验统计量 $U=\dfrac{\overline{x}-\mu_0}{\sigma/\sqrt{n}}$ 确定拒绝域的检验法称为 u 检验法。

【例 5-3】 某种子公司在销售胡萝卜种子的说明书中声称：用该种子生产的胡萝卜的平均长度为 11.5cm。某人买了 40 粒这种胡萝卜种子，种植后得到的胡萝卜长度的数据见表 5-2。若胡萝卜长度的标准差为 1.15cm，问在显著性水平 0.05 下，是否可以接受种子公司关于胡萝卜的平均长度的说明。

表 5-2 胡萝卜长度的数据

11.50	10.08	12.14	12.33	10.68	13.37	13.37	11.96	12.38	12.20
11.79	12.88	11.32	14.51	11.84	12.31	13.23	12.07	11.89	11.04
12.34	10.46	12.84	13.87	11.20	12.99	13.44	10.17	10.34	12.66
11.54	12.79	12.94	12.82	13.48	12.77	13.37	10.62	11.98	11.82

解：用 X 表示胡萝卜的长度，可以认为 $X\sim N(\mu,1.15^2)$。现在需要解答假设检验问题

$$H_0:\ \mu=11.5$$

调用 u 检验法函数 ztest，其调用格式如下：

```
[h, p, ci, u]=ztest(x, m, sigma, alpha, tail)
```

其中，输入参数 *x* 为样本数据向量，*m* 为待检验均值，*sigma* 为正态分布的标准差，*alpha* 为显著性水平（默认值为 0.05），*tail* 为检验的备择假设的标示值（*tail*=0 表示双侧检验，*tail*=1 表示右侧检验“>”，*tail*= –1 表示左侧检验“<”）；输出参数 *h* 为检验决策值（*h*=0 表示在显著性水平 *alpha* 下不能拒绝原假设，*h*=1 表示在显著性水平 *alpha* 下可以拒绝原假设），*p* 为拒绝原假设的最小显著性概率，*ci* 为真实均值 μ 的 1–*alpha* 置信区间，*u* 为检验统计量的值。

在 MATLAB 命令窗口中，将表 5-2 中的数据赋给列向量 x，然后运行程序代码

```
>> clear all;
x=[11.50,10.08,12.14,12.33,10.68,13.37,13.37,11.96,12.38,12.20,...
   11.79,12.88,11.32,14.51,11.84,12.31,13.23,12.07,11.89,11.04,...
   12.34,10.46,12.84,13.87,11.20,12.99,13.44,10.17,10.34,12.66,...
   11.54,12.79,12.94,12.82,13.48,12.77,13.37,10.62,11.98,11.82];
[h,sig]=ztest(x,11.5,1.15,0.05,0)
```

运行程序，输出如下：

```
h =
        1
sig =
   1.7154e-004
```



即在显著性水平 0.05 下拒绝原假设。

注意：① 在本例中，由已知计算出尾概率为 0.00017，因此在显著性水平 0.001 下，也要拒绝原假设，且这种拒绝犯错误的概率不超过 0.001。

② 在本例中，由于使用了双边检验，所以在显著性水平 0.05 下的检验结果只能拒绝种子公司关于胡萝卜的平均长度的说明，这种拒绝犯错误的概率为 0.05。

如果想进一步判断胡萝卜的平均长度是否大于 11.5cm，则需要解答单边检验问题。

2. σ^2 未知，关于 μ 的检验（t 检验法）

σ^2 未知，关于 μ 的检验主要有两类检验问题

- H_0： $\mu=\mu_0$； H_1： $\mu\neq\mu_0$。
- H_0： $\mu=\mu_0$； H_1： $\mu>\mu_0$（或 H_1： $\mu<\mu_0$）。

现就第一类检验问题检验法介绍如下。

设 $X_1,X_2,\cdots,X_N$ 是正态总体 $N(\mu,\sigma^2)$ 的样本，其中 σ^2 未知，μ 未知。用样本检验假设

$$H_0:\ \mu=\mu_0;\quad H_1:\ \mu\neq\mu_0$$

由于 σ^2 未知，故不能利用 $u=\dfrac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$ 来确定拒绝域了。注意到 S^2 是 σ^2 的无偏估计，现在用 S 来代替 σ，采用

$$T=\frac{\bar{X}-\mu_0}{S/\sqrt{n}}$$

作为统计量。当 H_0 为真时，由定理知

$$T=\frac{\bar{X}-\mu_0}{S/\sqrt{n}}\sim t(n-1)$$

于是对于给定的 α，由 t 分布表可查得临界值 $t_\alpha(n-1)$，使 $P\{|T|\geqslant t_\alpha(n-1)\}=\alpha$，即得拒绝域为 $\left(-\infty,-t_{\frac{\alpha}{2}}\right]\cup\left[t_{\frac{\alpha}{2}},+\infty\right)$。根据样本值算出检验统计量 T，将 $|t|$ 与 $t_{\frac{\alpha}{2}}(n-1)$ 比较，检验假设 H_0： $\mu=\mu_0$ 是否成立。

当 $|t|\geqslant t_\alpha(n-1)$ 时（即落入拒绝域），拒绝 H_0，接受 H_1。

当 $|t|<t_\alpha(n-1)$ 时（即落入相容域），接受 H_0。

对于第二类检验问题 H_0： $\mu=\mu_0$； H_1： $\mu>\mu_0$（或 H_1： $\mu<\mu_0$），当 σ^2 未知时，关于 μ 的单边检验步骤与上述内容类似，不同的是拒绝域的确定。请读者自行给出。

上述利用统计量得出的检验法称为 t 检验法。

【例 5-4】 设某次考试的学生成绩服从正态分布，从中随机抽取 36 位学生的成绩，算

得平均成绩为 66.5 分，标准差为 15 分。问在显著性水平 0.05 下，是否可以认为这次考试全体学生的平均成绩为 70 分？

解：由题意，待检验设为

$$H_0: \ \mu=70; \ H_1: \ \mu\neq 70$$

若 H_0 成立时，选取检验统计量

$$T=\frac{\bar{X}-\mu_0}{S/\sqrt{n}}$$

在显著性水平 $\alpha=0.05$ 下，查 t 分布表，得临界值

$$t_{\frac{\alpha}{2}}(n-1)=t_{0.025}(35)=2.0301$$

拒绝域为 $(-\infty,-2.0301]\cup[2.0301,+\infty)$。

检验统计量的值为

$$t=\frac{66.5-70}{15/\sqrt{36}}=-1.4\in(-2.0301,2.0301)$$

落在相容域中，故接受 H_0，即以可以认为在这次考试中全体学生的平均成绩为 70 分。

【例 5-5】 对于例 5-3 中的数据，问在显著性水平 α=0.05 下是否可以接受种子公司关于胡萝卜的平均长度的说明。

解：沿用例 5-3 中解答的符号，现在需要解答单边假设检验问题

$$H_0: \ \mu=11.5$$

由于题目中没有给出总体变量的标准差，所以现在面临的是在方差未知的情况下总体均值的双边假设检验问题。在 MATLAB 命令窗口中，将表 5-2 中的数据赋给列向量 x，然后运行程序代码

```
>> [h,sig]=ttest(x,11.5,0.05,0)
```

运行程序，输出如下：

```
h =
      1
sig =
   2.3808e-004
```

由结果可知，在显著性水平 α=0.05 下拒绝原假设。

注意：与例 5-3 中的尾概率 0.00017 相比较，这里的尾概率更大。这说明利用总体标准差的信息能够降低犯第一类错误的概率。

【例 5-6】 某车间用一台包装机包装葡萄糖，每袋葡萄糖的重量是一个随机变量，它服从正态分布。当机器正常时，其均值为 0.5kg，标准差为 0.015kg。某日开工后检验包装机是否正常，随机抽取 9 袋包装的糖，称得净重（单位：kg）为 0.497, 0.506, 0.518, 0.524, 0.498, 0.511, 0.52, 0.515, 0.512。问机器是否正常？



解：这是方差已知条件下正态分布均值的检验问题。注意到多数样本数据大于 0.5，故作单侧检验，检验假设为 H_0： $\mu=\mu_0=0.5$； H_1： $\mu>0.5$。注意，这里的原假设与备择假设是不相容的，但并非完全对立。这也是在实际应用中经常采用的检验命题的设定技巧。

其实现的 MATLAB 程序代码如下：

```
>> clear all;
x=[0.497,0.506,0.518,0.524,0.498,0.511,0.52,0.515,0.512];
[h,p,ci,u]=ztest(x,0.5,0.015,0.05,1)
```

运行程序，输出如下：

```
h =
     1
p =
    0.0124
ci =
    0.5030        Inf
u =
    2.2444
```

结果表明在显著性水平 $\alpha=0.05$ 下，可拒绝原假设，即认为包装机工作不正常，每袋葡萄糖的平均质量大于 0.5kg，由 *ci* 的值可知每袋葡萄糖的平均质量不低于 0.503kg 的置信水平为 0.95。

若忽视每袋葡萄糖质量的标准差已知的条件，则可调用函数 ttest 完成检验工作，其调用格式同 ztest 函数：

```
>> [h,p,ci,T]=ttest(x,0.5,0.05,1)
```

运行程序，输出如下：

```
h =
     1
p =
    0.0036
ci =
    0.5054        Inf
T =
    tstat: 3.5849
       df: 8
       sd: 0.0094
```

结果表明在 0.05 显著性水平下，*t* 检验亦拒绝原假设，即认为包装机工作不正常，每袋葡萄糖的平均质量大于 0.5kg。由 *p* 值可知，这个结论在显著性水平 $\alpha=0.01$ 下也是成立的。由 *ci* 的值可知每袋葡萄糖的平均质量不低于 0.5054kg 的置信水平为 0.99，结论错误的风险概率是 0.01。输出参数 *T* 报告检验统计量的观测值 *tstat*=3.5849，*t* 分布的自由度 *df*=8，对每袋葡萄糖质量标准的估计 *sd*=0.0094。

这里对例 5-3 稍作引申。生产商为确保产品投放市场后不出现较多的因质量指标不合格而引起的消费者投诉，在生产过程中实际的装袋质量往往大于向市场承诺的标准质量。在此例中，如果将袋装葡萄糖的平均质量为 0.5kg、标准差为 0.015kg 理解成是生产商对产品质量指标的承诺（而不是包装机的实际生产控制参数），则由每袋葡萄糖质量的样本标准差小于 0.01kg（更小于 0.015kg）可以认为，包装机的工作状态是平稳的。因此，样本均值大于 0.5kg 应是生产商确保质量指标承诺的体现。实际上，若以样本均值和样本标准差作为包装机的实际控制参数（估计），则可以推算出该生产商投放到市场上的袋装葡萄糖每袋质量大于 0.5kg 的比率，如下所示。

```
>> p=1-normcdf(0.5,mean(x),std(x))
```

运行程序，输出如下：

```
p =
    0.8840
```

即 88%的袋装葡萄糖的质量大于 0.5kg。

【例 5-7】 试用正态分布随机数函数生成一组随机数，并对该随机数进行均值假设检验。

解：假设先由 MATLAB 语句生成一组 400 个 $N(1,2^2)$ 的正态分布随机数，由于已知标准差为 2，可以引入假设 H_0： $\mu=1$，这样可以由下面的 MATLAB 语句进行检验，得出 $H=0$，故可以接受该假设。其实现的 MATLAB 程序代码如下：

```
>> r=normrnd(1,2,400,1);
>> [H,p,ci]=ztest(r,1,2,0.02)
```

运行程序，输出如下：

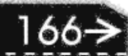
```
H =
     0
p =
    0.4034
ci =
    0.6838    1.1491
```

现在将假设设置为 H_0： $\mu=0.5$，则可以给出如下语句：

```
>> [H,p,ci]=ztest(r,0.5,2,0.02)
```

运行程序，输出如下：

```
H =
     1
p =
  3.1214e-005
ci =
    0.6838    1.1491
```



得出 $H=1$，表示应该拒绝 H_0 假设。若认为标准差未知，则可以采用 t 检验对假设 H_0： $\mu=1$ 进行检验，假设检验可以由下面的 MATLAB 语句直接得出。

```
>> [H,p,ci]=ttest(r,1,0.02)
```



运行程序，输出如下：

```
H =
     0
p =
    0.3756
ci =
    0.6964    1.1364
```

由于得出的 $H=0$，故表示可以接受该假设。

5.2.2 总体 $N(\mu,\sigma^2)$ 方差 σ^2 的检验

方差 σ^2 的假设检验在此主要讲解下列情形。

均值 μ 未知，方差 σ^2 的双边检验：

$$H_0:\ \sigma^2=\sigma_0^2;\ H_1:\ \sigma^2\neq\sigma_0^2$$

均值 μ 未知，方差 σ^2 的单边检验：

左边检验：H_0： $\sigma^2=\sigma_0^2$； H_1： $\sigma^2<\sigma_0^2$

H_0： $\sigma^2\geqslant\sigma_0^2$； H_1： $\sigma^2<\sigma_0^2$

右边检验：H_0： $\sigma^2=\sigma_0^2$； H_1： $\sigma^2>\sigma_0^2$

H_0： $\sigma^2\leqslant\sigma_0^2$； H_1： $\sigma^2<\sigma_0^2$

1．均值 μ 未知，方差 σ^2 的双边检验（χ^2 检验法）

设总体 $N(\mu,\sigma^2)$， μ,σ^2 均属未知， $X_1,X_2,\cdots,X_N$ 是样本。要求检验假设（显著性水平为 α）：H_0： $\sigma^2=\sigma_0^2$； H_1： $\sigma^2\neq\sigma_0^2$，其中 σ_0^2 为已知常数。

由于 S^2 是 σ^2 的无偏估计，当 H_0 为真时， S^2/σ_0^2 一般来说应在 1 附近摆动，而不应过分大于 1 或小于 1。若取检验统计量 $\chi^2=(n-1)S^2/\sigma_0^2$ ，则由定理知

$$\chi^2=(n-1)S^2/\sigma_0^2\sim\chi^2(n-1)$$

对显著性水平 α ，拒绝域具有形式

$$\chi^2\leqslant\lambda_1 \text{ 或 } \chi^2\geqslant\lambda_2$$

使

$$P\{\chi^2\leqslant\lambda_1\}+P\{\chi^2\geqslant\lambda_2\}=\alpha$$

为计算方便，习惯上取

$$P\{\chi^2\leqslant\lambda_1\}=\frac{\alpha}{2},\quad P\{\chi^2\geqslant\lambda_2\}=\frac{\alpha}{2}$$

故得 $\lambda_1=\chi^2_{1-\frac{\alpha}{2}}(n-1)$，$\lambda_2=\chi^2_{\frac{\alpha}{2}}(n-1)$，于是拒绝域为 $\left[0,\chi^2_{1-\frac{\alpha}{2}}(n-1)\right]\cup\left[\chi^2_{\frac{\alpha}{2}}(n-1),+\infty\right)$，相容域为 $\left(\chi^2_{1-\frac{\alpha}{2}}(n-1),\chi^2_{\frac{\alpha}{2}}(n-1)\right)$。

上述检验法由于检验统计量为 χ^2 且符合 χ^2 分布，故称为 χ^2 检验法。

【例 5-8】 某工厂生产铜丝，工艺改进后产量提高。现从产品中抽出 10 根检查拉断力，得数据为 572, 578, 570, 568, 570, 572, 570, 572, 596, 584。从以往资料和技术标准上知拉断力的方差 σ^2 为 64 时合格，否则认为是不合格的。问工艺改进后铜丝产品能否认为是合格的？

解：设 X 为铜丝的拉断力，根据经验知 $X\sim N(\mu,\sigma^2)$。由题意知，需检验假设

$$H_0:\ \sigma^2=\sigma_0^2=64,\quad H_1:\ \sigma^2\neq 64$$

由样本值算出 $\bar{x}=575.2$

$$\sum_{i=1}^{10}(x_i-\bar{x})^2=\sum_{i=1}^{10}x_i^2-n\bar{x}^2=3309232-3308550.4=681.6$$

$$\chi^2=\sum_{i=1}^{10}(x_i-\bar{x})^2\Big/\sigma_0^2=681.6/64=10.65$$

取 $\alpha=0.05$，$n=10$，查自由度为 9 的 χ^2 分布表，知 $\chi^2_{0.025}(9)=19.0$，$\chi^2_{0.975}(9)=2.7$，得拒绝域为 $[0,2.70]\cup[19.0,+\infty]$，相容域为 $(2.70,19.0)$。由于 $\chi^2=10.65\in(2.70,19.0)$ 落在相容域中，故接受 H_0，即在显著性水平 $\alpha=0.05$ 下，工艺改进后铜丝产品是合格的。

现取 $\alpha=0.01$，检验统计量选为 $\chi^2=\dfrac{(n-1)S^2}{\sigma^2}\sim\chi^2(n-1)$，由 χ^2 分布表知 $\chi^2_{1-\alpha}(n-1)=\chi^2_{0.99}(9)=2.088$，得 H_0 的拒绝域为 $\chi_0^2\leqslant\chi^2_{1-\alpha}(n-1)$，由样本算出 $\chi_0^2=218.1/14^2=1.113<2.088$（在拒绝域内），故拒绝 H_0，接受 H_1，即提纯后的样本高度更整齐。

2．均值 μ 未知，方差 σ^2 的单边检验（χ^2 检验法）

在此讲解正态总体 $N(\mu,\sigma^2)$，μ,σ^2 均未知，假设为

$$H_0:\ \sigma^2\leqslant\sigma_0^2,\ \text{其中}\ \sigma_0^2\ \text{为已知常数；}\ H_1:\ \sigma^2>\sigma_0^2$$

的检验问题。顺便指出，这种检验在实际中很有应用价值，生产中为了了解加工精度有无变化，进行抽样，如算得样本方差 S^2 比原来的方差 σ_0^2 大，这时可检验假设 H_0：$\sigma^2\leqslant\sigma_0^2$。经过检验，如果能否定 H_0，说明精度降低了，需停产检查原因；否则，精度没有降低。

此问题分析如下：设 $X_1,X_2,\cdots,X_N$ 是来自总体 $N(\mu,\sigma^2)$ 的样本，若 $S^2\big/\sigma_0^2$ 很大，则有理由否定假设 H_0：$\sigma^2\leqslant\sigma_0^2$，否则，可以接受这个假设。

在假设 H_0 为真的情况下，S^2/σ_0^2 的概率分布并不能算出来，但有（如前所述）

$$\frac{(n-1)S^2}{\sigma_0^2}\sim\chi^2(n-1)$$

于是，对显著性水平 α，有临界值 $\lambda=\chi^2_\alpha(n-1)$，使

$$P\left\{\frac{(n-1)S^2}{\sigma_0^2}\geqslant\lambda\right\}=\alpha$$

但由于σ^2未知，故$\frac{(n-1)S^2}{\sigma_0^2}$算不出来。在假设$H_0$：$\sigma^2\leqslant\sigma_0^2$下，有



$$\frac{(n-1)S^2}{\sigma_0^2}\leqslant\frac{(n-1)S^2}{\sigma^2}$$

因此

$$P\left\{\frac{(n-1)S^2}{\sigma_0^2}\geqslant\lambda\right\}\leqslant P\left\{\frac{(n-1)S^2}{\sigma^2}\geqslant\lambda\right\}=\alpha$$

这就表明，事件$\left\{\frac{(n-1)S^2}{\sigma_0^2}\right\}$更是一个小概率事件，从而有

$$H_0：\sigma^2\leqslant\sigma_0^2；H_1：\sigma^2>\sigma_0^2，\sigma_0^2\text{为已知常数}$$

检验统计量$\chi^2=\frac{(n-1)S^2}{\sigma_0^2}$，对显著性水平$\alpha$，临界值$\lambda=\chi_\alpha^2(n-1)$，拒绝域为$[\chi_\alpha^2(n-1)$，$+\infty]$，当检验统计量的值$\chi^2\geqslant\chi_\alpha^2(n-1)$时，拒绝$H_0$；否则，接受$H_0$。

其他情形方差σ^2的单边检验也有类似的讲解，这里从略。

【例 5-9】 电工器材厂生产一批熔丝，抽取 10 根试验其熔断时间，结果为 42, 65, 75, 78, 71, 59, 57, 68, 54, 55。
已知熔断时间服从正态分布。问是否可认为整批熔丝的熔断时间的方差不大于 80？取显著性水平$\alpha=0.05$。

解：由题意知，待检验假设为

$$H_0：\sigma^2\leqslant 80，H_1：\sigma^2>80$$

此检验为右边检验。对显著性水平α，临界值为$\chi_\alpha^2(n-1)=\chi_{0.05}^2(9)=16.919$，拒绝域为$[16.919,+\infty)$。

由样本值可知$\overline{x}=62.4$，$\sum_{i=1}^{10}(x_i-\overline{x})^2=1096.4$，检验统计量的值

$$\chi^2=\frac{(n-1)S^2}{\sigma_0^2}=\frac{\sum_{i=1}^{10}(x_i-\overline{x})^2}{\sigma_0^2}=\frac{1096.4}{80}$$

故接受H_0，即在显著性水平$\alpha=0.05$下，可以认为整批熔丝的熔断时间的方差不大于 80。

5.3 两正态总体参数的假设检验

上面讲解了单个正态总体参数的显著性检验，它是把样本统计量的观察值与原假设所提供的总体参数作比较，这种检验要求事先能提出合理的假设值，并对参数有某种意义的备择

值，但在实际工作中很难做到这一步，因而限制了这种方法在实际中的应用。实际中常常选择两个样本，一个作为处理，另一个作为对照。在两个样本间作比较，如比较两种处理之间的差异，两种实验方法或两种药物的疗效等，判断它们之间是否存在足够显著的差异。或者说，判断它们之间的差异能否用偶然性解释，当不能用偶然性解释时，则认为它们之间存在足够显著的差异，从而推断两个样本来自不同的总体。

5.3.1 方差未知但相等时两个正态总体均值的检验

设有两个独立的正态总体 $X \sim N(\mu_1,\sigma_1^2)$，$Y \sim N(\mu_2,\sigma_2^2)$，$X_1,X_2,\cdots,X_{n_1}$ 与 $Y_1,Y_2,\cdots,Y_{n_2}$ 分别是 X 和 Y 的样本，$\bar{X}$，$\bar{Y}$，S_1^2，S_2^2 是相应的样本均值和样本方差。常见的关于均值的假设检验如下。

- H_0：$\mu_1=\mu_2$；H_1：$\mu_1 \neq \mu_2$（称为双边检验，H_1 可略而不写）。
- H_0：$\mu_1=\mu_2$；H_1：$\mu_1 > \mu_2$ 或 H_1：$\mu_1 \leqslant \mu_2$；H_1：$\mu_1 > \mu_2$（称为右边检验）。
- H_0：$\mu_1=\mu_2$；H_1：$\mu_1 < \mu_2$ 或 H_1：$\mu_1 \geqslant \mu_2$；H_1：$\mu_1 < \mu_2$（称为左边检验）。

以 σ_1^2，σ_2^2 未知，但 $\sigma_1^2=\sigma_2^2$ 为例，讲解检验假设 H_0：$\mu_1=\mu_2$；H_1：$\mu_1 \neq \mu_2$，即得到检验统计量为

$$T=\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{S_W\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}} \sim t(n_1+n_2-2)$$

其中，$S_W^2=\dfrac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$，$S_1^2=\dfrac{1}{n_1-1}\sum\limits_{i=1}^{n_1}(X_i-\bar{X})^2$，$S_2^2=\dfrac{1}{n_2-1}\sum\limits_{i=1}^{n_2}(Y_i-\bar{Y})^2$。

当 H_0 成立时，检验统计量

$$T=\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{S_W\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}} \sim t(n_1+n_2-2)$$

于是，对给定的显著性水平 α，查 t 分布表，取临界值 $t_{\frac{\alpha}{2}}(n_1+n_2-2)$，由 $P\left\{|T| \geqslant t_{\frac{\alpha}{2}}(n_1+n_2-2)\right\}=\alpha$ 得 H_0 的拒绝域为

$$|T_0| \geqslant t_{\frac{\alpha}{2}}(n_1+n_2-2)$$

【例 5-10】 设甲、乙两煤矿出煤的含灰率（单位：%）都服从正态分布，即 $X \sim N(\mu_1,7.5)$，$Y \sim N(\mu_2,2.6)$。为检验两煤矿的煤含灰率有无显著性差异，从两煤矿中各取样若干份，分析结果如下。

甲矿：24.3, 20.8, 23.7, 21.3, 17.4

乙矿：18.2, 16.9, 20.2, 16.7

试在显著性水平 $\alpha=0.05$ 下，检验“含灰率无差异”这个假设。

分析：检验假设为

$$H_0：\mu_1=\mu_2；\ H_1：\mu_1 \neq \mu_2$$

取检验统计量 $\bar{X}-\bar{Y}$，由于 σ_1^2，σ_2^2 均已知，统计量规范化为 $U=\dfrac{\bar{X}-\bar{Y}}{\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}}\sim N(0,1)$，检验准则是 $P\{|U|\geqslant\delta\}\leqslant\alpha$，即拒绝域为 $|U|\geqslant\delta$。



其实现的 MATLAB 程序代码如下：

```
>> clear all;
x=[24.3,20.8,23.7,21.3,17.4];
y=[18.2,16.9,20.2,16.7];
alpha=0.05;  %设定显著性水平
U=(mean(x)-mean(y))/sqrt(7.5/5+2.6/4);  %计算检验统计量的观测值
DETA=norminv((1-alpha/2),0,1);  %求拒绝域的临界值
p=1-normcdf(U,0,1);  %求拒绝原假设的最小显著性概率
if abs(U)>DETA  %决策,拒绝原假设，则返回 h=1；否则返回 h=0
    h=1;
else
    h=0;
end
alpha,h,p,U,DETA
```

运行程序，输出如下：

```
alpha =
   0.0500
h =
     1
p =
    0.0085
U =
   2.3870
DETA =
    1.9600
```

结果表明，在显著性水平 $\alpha=0.05$ 下，认为甲矿含灰率与乙矿含灰率有显著性差异。

若注意到含灰率数据的均值甲矿明显大于乙矿，进行单侧检验更为恰当，检验假设可表示为

$$H_0:\ \mu_1=\mu_2;\quad H_1:\ \mu_1>\mu_2$$

此时，检验准则是 $P\{|U|\geqslant\delta\}\leqslant\alpha$，即拒绝域为 $U\geqslant\delta$。相应的数据处理过程只需在上述 MATLAB 指令集中，将语句

```
>>DETA=norminv((1-alpha/2),0,1)
```

修改为

```
>> DETA=norminv((1-alpha),0,1)
```

输出如下：

```
DETA =     1.6449
```

即可，此时 DETA = 1.6449，其他计算结果不变。相应的检验结论是：在显著性水平 $\alpha=0.05$ 下，认为甲矿含灰率显著地大于乙矿含灰率。由 p 值可知，这个结论在显著性水平 $\alpha=0.01$ 下也是成立的。

MATLAB 给出了方差未知但相等的条件下，用于两个正态变量均值差的检验函数 ttest2，其使用方法与函数 ttest 类似。

【例 5-11】 在平炉上进行一项试验，以确定改变操作方法的建议是否会增加钢的产率。试验是在同一只平炉上进行的。每炼一炉钢时除操作方法外，其他条件都尽可能做到相同。先用标准操作方法炼一炉，然后用建议的新操作方法炼一炉，以后交替进行，各炼 10 炉，其产率分别如下：

标准操作方法：78.1, 72.4, 76.2, 74.3, 77.4, 78.4, 76.0, 75.5, 76.7, 77.3

新操作方法：79.1, 81.0, 77.3, 79.1, 80.0, 79.1, 79.1, 77.3, 80.2, 82.1

设这两个样本相互独立，且分别来自正态总体 $N(\mu_1,\sigma^2)$ 和 $N(\mu_2,\sigma^2)$，μ_1,μ_2,σ^2 均未知。问建议的新操作方法能否提高产率（取 $\alpha=0.05$）？

其实现的 MATLAB 程序代码如下：

```
>> clear all;
X=[78.1, 72.4, 76.2, 74.3, 77.4, 78.4, 76.0, 75.5, 76.7, 77.3];
Y=[79.1, 81.0, 77.3, 79.1, 80.0, 79.1, 79.1, 77.3, 80.2, 82.1];
[h,sig,ci]=ttest2(X,Y,0.05,-1)
```

运行程序，输出如下：

```
h =       1
sig =   2.1759e-004
ci =
        -Inf    -1.9083
```

h=1 表示在显著性水平 $\alpha=0.05$ 下应该不接受原假设，*sig*=2.1759e–004 表明两个总体均值相等的概率很小，因此认为建议的新操作方法提高了产率，比标准操作方法好。

5.3.2 两个正态总体方差齐性（相等）的检验

5.3.1 节讲解了两个正态总体方差未知但相等时，总体均值的检验。然而又怎样得出方差相等的结论呢？这需要对方差本身进行检验。只有通过检验接受方差这一假设，才能进行上面的两个正态总体的均值检验。

设两个正态总体 $X\sim N(\mu_1,\sigma_1^2)$，$Y\sim N(\mu_2,\sigma_2^2)$，且相互独立，$X_1,X_2,\cdots,X_{n_1}$ 与 $Y_1,Y_2,\cdots,Y_{n_2}$ 分别是 X 和 Y 的样本。下面仅就 μ_1，μ_2 未知时，讲解假设检验

$$H_0:\ \sigma_1^2=\sigma_2^2;\ H_1:\ \sigma_1^2\neq\sigma_2^2$$

由前面的学习知，统计量

$$F=\frac{S_1^2\sigma_2^2}{S_2^2\sigma_1^2}\sim F(n_1-1,n_2-1)$$

当 H_0 为真时，则统计量



$$F=\frac{S_1^2}{S_2^2}\sim F(n_1-1,n_2-1)$$

于是查 F 分布表（在相关资料中查阅 F 分布表），取临界值 $F_{\frac{\alpha}{2}}(n_1-1,n_2-1)$ 和 $F_{1-\frac{\alpha}{2}}(n_1-1,n_2-1)$，使

$$P\left\{F\leqslant F_{1-\frac{\alpha}{2}}(n_1-1,n_2-1)\right\}\cup F\geqslant F_{\frac{\alpha}{2}}(n_1-1,n_2-1)=\alpha$$

得到 H_0 的拒绝域为

$$F_0\geqslant F_{\frac{\alpha}{2}}(n_1-1,n_2-1)\text{ 或 }F_0\leqslant F_{1-\frac{\alpha}{2}}(n_1-1,n_2-1)$$

【例 5-12】 用两种方法研究冰的潜热，样本均取自–0.72℃的冰。用方法 A 做，取样本容量 $n1$=13；用方法 B 做，取样本容量 $n2$=8，测量每克冰从–0.72℃变 0℃的水。其中，热量的变化数据见表 5-3。

表 5-3 热量的变化数据

方法 A	79.98	80.04	80.02	80.04	80.03	80.04	80.03	79.97	80.05	80.03	80.02	80.00	80.02
方法 B	80.02	79.94	79.97	79.98	79.97	80.03	79.95	79.97					

假设两种方法测得数据总体都服从正态分布。这两种研究方法有无显著性差异（$\alpha=0.05$）？两组数据的方差是否具有齐性？

解：检验 H_0：$\sigma_1^2=\sigma_2^2$；H_1：$\sigma_1^2\neq\sigma_2^2$

选取统计量

$$F=\frac{S_1^2}{S_2^2}\sim F(n_1-1,n_2-1)$$

则 H_0 为真时，拒绝域为

$$F_0\geqslant F_{\frac{\alpha}{2}}(n_1-1,n_2-1)\text{ 或 }F_0\leqslant F_{1-\frac{\alpha}{2}}(n_1-1,n_2-1)$$

计算有关数据

$$\overline{x}=80.02,\quad S_1^2=5.75\times10^{-4},\quad \overline{y}=79.98,\quad S_2^2=9.86\times10^{-4},\quad F=\frac{S_1^2}{S_2^2}=0.5832$$

又查 F 分布表，得

$$F_{\frac{\alpha}{2}}(n_1-1,n_2-1)=F_{0.025}(12,7)=3.61$$

而

$$F_{1-\frac{\alpha}{2}}(n_1-1,n_2-1)=F_{0.975}(12,7)=\frac{1}{F_{0.025}(12,7)}=\frac{1}{3.61}=0.277$$

因 $0.277 < 0.5832 < 3.61$，即 F_0 的值在 H_0 的接受域内，故接受 H_0，说明两测试总体的方差相等。故两组数据的方差具有齐性。

5.4 非正态总体参数的假设检验

设总体 X 服从某种非正态分布，其分布函数 $F(X;\theta)$ 中含有未知参数 θ，即其概率函数 $p(x;\theta)$ 或概率密度函数 $f(x;\theta)$ 中含有未知参数 θ。那么其数学期望 $E(X)=\mu(\theta)$，方差 $D(X)=\sigma^2(\theta)$ 都是未知参数 θ 的函数。从总体中抽取大容量简单随机样本 $X_1,X_2,\cdots,X_n$ 为依据，来检验原假设 H_0： $\theta=\theta_0$。

由于简单随机样本 $X_1,X_2,\cdots,X_n$ 相互独立，且与总体 X 服从相同的分布，因而，在原假设 H_0： $\theta=\theta_0$ 成立的条件下，由“独立同分布的林德伯格-列维中心极限定理”知：当样本容量 n 充分大时（一般 n>50），统计量

$$U=\frac{\sum_{i=1}^{n}X_i-nE(X)}{\sqrt{nD(X)}}=\frac{\sum_{i=1}^{n}X_i-n\mu(\theta_0)}{\sqrt{n\sigma^2(\theta_0)}}=\frac{\bar{X}-\mu(\theta_0)}{\frac{\sigma(\theta_0)}{\sqrt{n}}}$$

近似服从标准正态分布 $N(0,1)$。因此，在给定的显著性水平 α 下，有

$$P\left\{|U|>Z_{\frac{\alpha}{2}}\right\}\approx\alpha(\text{双侧})；\ P\{U>Z_\alpha\}\approx\alpha(\text{单侧}),\ P\{U<-Z_\alpha\}\approx\alpha(\text{单侧})$$

【例 5-13】 某厂产品的优质品率一直保持在 40%，近期技监部门来厂抽查，共抽查了 12 件产品，其中优质品为 5 件，在显著性水平 $\alpha=0.05$ 下能否认为其优质率仍保持在 40%？

分析：设 X 表示检查一个产品时优质品的个数，则 $X\sim b(1,p)$。检验问题为

$$H_0:\ p=0.4；\ H_1:\ p\neq 0.4$$

这是一个双边检验问题。当 H_0 为真时，检验统计量 $T=\sum_{i=1}^{n}X_i\sim b(12,0.4)$，拒绝域为 $T\leqslant c_1$ 或 $T\geqslant c_2$ $(c_1<c_2)$。其中，临界值 c_1 是使 $P\{T\leqslant c_1\}\leqslant 0.025$ 成立的最大整数，c_2 是使 $P\{T\geqslant c_2\}\leqslant 0.025$ 成立的最小整数。

其实现的 MATLAB 程序代码如下：

```
>>clear all;
T=5;                          %检验统计量的观测值
alpha=0.025;                  %显著性水平
p=binocdf(0:12,12,0.4);       %为确定拒绝域临界值，计算 T 的累积概率
for byk=1:7                   %求拒绝域临界值
    if p(byk)<=alpha&p(byk+1)>=alpha
        c1=byk-1;
    end
    if(1-p(byk+6))>alpha&(1-p(byk+7))<=alpha
        c2=byk+7;
    end
```

```
end
if T<=c1|T>=c2      %检验决策,h=1(0)拒绝(接受)原假设
      h=1
else
      h=0
end
c=[c1,c2]           %输出拒绝域临界值
```



运行程序，输出如下：

```
h =
      0
c =
       1      9
```

上述计算表明，当显著性水平 $\alpha=0.05$ 时，由于 $P\{T\leqslant 1\}<0.025$ 而 $P\{T\leqslant 2\}>0.025$，故拒绝域的左侧临界值 $c_1=1$；又因为 $P\{T\geqslant 8\}>0.025$ 而 $P\{T\geqslant 9\}<0.025$，故拒绝域的右侧临界值 $c_2=9$。于是，H_0 的拒绝域为 $T\leqslant 1$ 或 $T\geqslant 9$。检验统计量的观测值 $T=5$ 未落入拒绝域，因而在显著性水平 $\alpha=0.05$ 下，认为该厂的优质品率无明显变化。

【例 5-14】 从随机抽取的 467 名男性中发现有 8 名色盲，而 433 名女性中发现有 1 人色盲，在显著性水平 $\alpha=0.01$ 下能否认为女性色盲的比例比男性低？

分析：设男性色盲的比例为 p_1，女性色盲的比例为 p_2，那么要检验的假设为 H_0：$p_1\geqslant p_2$；H_1：$p_1<p_2$。

其实现的 MATLAB 程序代码如下：

```
>> clear all;
alpha=0.01;                          %显著性水平
ESTp1=8/467;
ESTp2=1/433;
ESTp=(8+1)/(467+433);
U=(ESTp1-ESTp2)/sqrt((1/467+1/433)*ESTp*(1-ESTp)) %检验统计量的观测值
c=norminv(alpha,0,1)                 %求拒绝域的临界值
if U<=c                              %检验决策，h=1(0)拒绝（接受）原假设
      h=1
else
      h=0
end
```

运行程序，输出如下：

```
U =
   2.2328
c =
   -2.3263
h =
      0
```

结果表明，在显著性水平$\alpha = 0.01$下不能拒绝原假设，即可以认为女性色盲的比例比男性低。

5.5 变量分布形态的检验

通过前几节的讲解，已经了解了假设检验的基本思想，并讲解了当分布形式已知时关于其中未知参数的假设检验问题。然而，可能遇到这样的情形，如例 5-10 中，认为标准方法下的钢的产率服从正态分布通常是合理的，但是新操作方法下钢的产率是否仍服从正态分布是需要斟酌的，因为影响钢的产率的条件发生了改变。因此在例 5-10 问题的分析中，更为严谨的思考应当包括识别新操作方法下钢的产率是否为某个正态变量。此类问题通常称为变量分布形态的检验，属于非参数检验问题。本节讲解非参数检验的几个基本方法及其应用。

5.5.1 χ^2 拟合优度检验

例如，某公司雇用 200 名员工，男性和女性员工人数分别为：男性 150 名，女性 50 名，该公司被指控在雇用员工时有性别歧视。要调查这项指控，需要考虑在没有歧视的情况下，人们期望这两种性别的员工人数。换句话说，把期望的频率与实际观测的频率进行比较，就产生了拟合优度检验问题，即如果观测频率与期望频率拟合优度较好，则可以得出结论：在给定的显著性水平下，公司没有歧视。该检验称为χ^2拟合优度检验。

为了介绍拟合优度检验的原理，来分析一下性别歧视问题。需要确定如果没有歧视，人们期望雇用每一性别的人数是多少。一种方法是考虑全体雇员中男女性别的比例——分别为60%和 40%。这意味着期望该公司雇用 120 名男性和 80 名女性，见表 5-4。

表 5-4　某公司员工性别期望表

性　别	男　性	女　性
观测频率	150	50
期望频率	120	80

当然，如果在每一种性别中观测频率和期望频率没有差别，那么这足以证明不存在歧视。如果存在差别（如这里的情况），那么提出问题的差别是由偶然性引起的或是差别太大而不仅仅是由偶然性引起的。因此，需要构造基于观测频率和期望频率之间的差别的统计量。

卡尔·皮尔逊最先提出了统计量χ^2可作为度量经验分布与假设分布之间的差异来检验H_0是否成立。χ^2检验要求假设H_0中的总体分布$F_0(x)$的形式及其参数必须是已知的，但实际中参数往往是未知的。通常，需要先用极大似然估计法估计出$F_0(x)$中的参数，再作检验。

设总体是 m 个可能的离散型随机变量，不失一般性，设 X 的可能值是$1,2,\cdots,m$，记它取值为 i 的概率为 p_i，即

$$P\{X = i\} = p_i, i = 1,2,\cdots,m\text{，显然有}\sum_{i=1}^{m} p_i = 1$$



设 $X_1,X_2,\cdots,X_n$ 是从总体 X 中抽取的简单随机样本，$x_1,x_2,\cdots,x_n$ 是样本观察值。记 n_i 为 $x_1,x_2,\cdots,x_n$ 中取值为 i 的个数，即样本中出现事件 $(X=i)$ 的频数。由大数定律知道，频率是概率的反映。如果总体的概率分布的确是 $p_{10},p_{20},\cdots,p_{m0}$，那么，当观察个数 n 越来越大时，频率 $\frac{n_i}{n}$ 与 p_{i0} 之间的差异将越来越小，且 $\chi^2=\sum_{i=1}^{m}\left(\frac{n_i}{n}-p_{i0}\right)^2$ 也较小。根据这一思想，卡尔·皮尔逊提出了运用统计量

$$\chi^2=\sum_{i=1}^{m}\frac{(n_i-np_{i0})^2}{np_{i0}} \tag{5-1}$$

来反映它们的差异程度，式（5-1）也称为卡尔·皮尔逊统计量。

定理 5-1（K.Pearsn 定理） 当 $p_{10},p_{20},\cdots,p_{m0}$ 是总体的真实概率分布时，由式（5-1）所定义的统计量 χ^2 渐进服从自由度为 $m-1$ 的 χ^2 分布，即

$$\chi^2=\sum_{i=1}^{m}\frac{(n_i-np_{i0})^2}{np_{i0}}\approx\chi^2(m-1)$$

根据这个定理，当样本容量足够大时，就近似地认为统计量 $\chi^2=\sum_{i=1}^{m}\frac{(n_i-np_{i0})^2}{np_{i0}}\sim\chi^2(m-1)$，此时卡尔·皮尔逊统计量的值一般比较小，因此，当假设 H_0：$p_i=p_{i0}$；H_1：$p_i\neq p_{i0}$ $(i=1,2,\cdots,m)$，其中 p_{i0} 是已知数。只要算出观察值 $\chi^2=\sum_{i=1}^{m}\frac{(n_i-np_{i0})^2}{np_{i0}}$，对于给定的显著性水平 $0<\alpha<1$，由 χ^2 分布表求出常数 $\chi_\alpha^2(m-1)$，使

$$P\{\chi^2\geqslant\chi_\alpha^2(m-1)\}=\alpha$$

如果 $\chi_0^2\geqslant\chi_\alpha^2(m-1)$，则拒绝 H_0，即认为总体的分布与假设 H_0 中的分布有显著差异；若 $\chi_0^2<\chi_\alpha^2(m-1)$，则接受 H_0，即认为总体的分布与假设 H_0 中的分布无显著差异。

现用 χ^2 检验法来检验性别有无歧视问题。假设人们期望员工性别比例为 6∶4，即男性为 120 人，女性为 80 人。

假设 H_0：公司对员工的性别无歧视，计算得

$$\chi_0^2=\frac{(150-120)^2}{120}+\frac{(50-80)^2}{80}=7.5+11.5=18.75$$

对 $\alpha=0.05$，查 χ^2 分布表得 $\chi_{0.05}^2(2-1)=3.841$，$\chi_0^2>\chi_{0.05}^2(1)$，故拒绝 H_0，说明有明显差异。这对性别歧视指控提供了证据。

下面举例说明 χ^2 拟合优度检验法的应用。

【例 5-15】 表 5-5 中数据是 200 个零件的直径 X。

表 5-5　200 个零件的直径数据　　（单位：cm）

直　径	2.25	2.35	2.45	2.55	2.65	2.75	2.85	2.95
频　数	3	4	5	11	12	17	19	26
直　径	3.05	3.15	3.25	3.35	3.45	3.55	3.65	3.75
频　数	24	22	19	13	13	7	3	2

能否验证直径X服从正态分布？

分析：依题意，检验的假设是H_0：零件直径 X 服从正态分布$N(\mu,\sigma^2)$。其中，参数μ,σ^2均未知。因此，首先要求出参数μ,σ^2的极大似然估计：

$$\hat{\mu}_{MLE}=\frac{1}{n}\sum_{i=1}^{k}f_i x_i$$（分组数据的样本均值）

$$\hat{\sigma}^2_{MLE}=\frac{1}{n}\sum_{i=1}^{k}f_i x_i^2-\mu^2_{MLE}$$（分组数据的样本方差）

然后按照以下步骤进行χ^2拟合优度检验。

其实现的 MATLAB 程序代码如下：

① 输入原始数据，并求分布参数的极大似然估计。

```
>> clear all;
x=[2.25,2.35,2.45,2.55,2.65,2.75,2.85,2.95,3.05,3.15,3.25,3.35,3.45,3.55,3.65,3.75];
f=[3 4 5 11 12 17 19 26 24 22 19 13 13 7 3 2];
n=sum(f);
MU=sum(f.*x)./n
SIGMA=sqrt(sum(f.*(x.^2))./n-MU.^2)
```

运行程序，输出如下：

```
MU =      3.0087
SIGMA =      0.3217
```

根据计算结果，检验的原假设修正为H_0： $X\sim N(3.009,0.3210^2)$。

② 样本数据分组。

题目给出的数据已是分组数据，共分为 16 组，且每组的频数已经统计出。但是，前 3 组数据和后 3 组数据的频数偏小，故分别将前、后 3 组数据进行合并，这样可得 12 组数据。这 12 组数据所属的数据组的区间边界值如下。

```
a=[];
for k=1:11
    aa=(x(2+k)+x(3+k))/2;  %小区间边界点取相邻两个数据的中点
    a=[a,aa];
end
a=[-inf,a,inf]'  %由于正态变量在整个数轴上取值,最小边界点为-∞，最大边界点为+∞
```

运行程序，输出如下：

```
a =
    -Inf
   2.5000
   2.6000
   2.7000
   2.8000
   2.9000
```



```
    3.0000
    3.1000
    3.2000
    3.3000
    3.4000
    3.5000
       Inf
```

③ 统计经验频数。

题目已经给出经验频数，只需分别合并前、后 3 组的频数。

```
>> f=[f(1)+f(2)+f(3),f(4:13),f(14)+f(15)+f(16)]'
```

运行程序，输出如下：

```
f =
    12
    11
    12
    17
    19
    26
    24
    22
    19
    13
    13
    12
```

④ 计算理论频数。

```
>> PEST=[];
for i=1:12
    pp=normcdf(a(i+1),MU,SIGMA)-normcdf(a(i),MU,SIGMA);
    PEST=[PEST,pp];
end
THEF=n*PEST'
```

运行程序，输出如下：

```
THEF =
   11.3794
    9.0111
   13.3332
   17.9247
   21.8947
   24.2992
   24.5027
```

```
22.4494
18.6880
14.1347
 9.7136
12.6692
```

⑤计算检验统计量的观测值。

```
>> CHI2EST=sum((f-THEF).^2./THEF)
```

运行程序，输出如下：

```
CHI2EST =
    2.4184
```

⑥ 检验决策。

```
k=12;
r=2;
alpha=0.05;
df=k-r-1;
REFCR=chi2inv(1-alpha,df);          %拒绝域临界值
p=1-chi2cdf(CHI2EST,df);            %检验的 p 值
if CHI2EST>REFCR
    h=1;
else
    h=0;
end
alpha,h,p
stat=[k,r,CHI2EST,REFCR]
```

运行程序，输出如下：

```
alpha =      0.0500
h =        0
p =      0.9830
stat =
   12.0000      2.0000      2.4184     16.9190
```

计算结果表明，在显著性水平 $\alpha=0.05$ 下，h=0 保留原假设 H_0，即 χ^2 拟合优度检验认为零件直径 $X\sim N(3.009,0.1030)$。最小显著性概率 p=0.9823 表明，当前样本数据下不能拒绝原假设 H_0 的置信水平高达 0.98。

【例 5-16】 在 20 天内，从维尼纶正常生产时的生产报表中看到的维尼纶纤度（纤维的粗细程度的一种度量）的情况，有如下 100 个数据。

```
1.36,1.49,1.43,1.41,1.37,1.40,1.32,1.43,1.47,1.39,
1.41,1.36,1.40,1.34,1.42,1.42,1.45,1.35,1.42,1.39,
1.44,1.42,1.39,1.42,1.42,1.30,1.34,1.42,1.37,1.36,
```

1.37,1.34,1.37,1.37,1.44,1.45,1.32,1.48,1.40,1.45,
1.39,1.46,1.39,1.53,1.36,1.48,1.40,1.39,1.38,1.40,
1.36,1.45,1.50,1.43,1.38,1.43,1.41,1.48,1.39,1.45,
1.37,1.37,1.39,1.45,1.31,1.41,1.44,1.44,1.42,1.42,
1.35,1.36,1.39,1.40,1.38,1.35,1.42,1.43,1.42,1.42,
1.42,1.40,1.41,1.37,1.46,1.36,1.37,1.27,1.37,1.38,
1.42,1.34,1.43,1.42,1.41,1.41,1.44,1.48,1.55,1.37.



正常情况下，维尼纶纤度服从正态分布。试根据这 100 个样本数据在显著性水平 $\alpha=0.10$ 下验证生产是正常的。

分析：这是一个正态拟合问题。检验的原假设是 H_0：维尼纶纤度 X 服从正态分布 $N(\mu,\sigma^2)$。其中，参数 μ,σ^2 均未知。

其实现的 MATLAB 程序代码如下：

① 输入原始数据，进行未知参数的极大似然估计。

```
>> clear all;
load data. mat;   %预先编写数据文件 data.mat,并存放在当前工作路径下
n=length(data);
[MU,SIGMA]=normfit(data)
```

运行程序，输出如下：

```
MU =       1.4338
SIGMA =       0.3043
```

于是，检验假设修正为 H_0： $X\sim N(1.4042,0.0178^2)$

② 样本数据分组。

```
>> [f,med]=hist(data);
F_MED=[f',med']
```

运行程序，输出如下：

```
F_MED =
    99.0000      1.4270
          0       1.7410
          0       2.0550
          0       2.3690
          0       2.6830
          0       2.9970
          0       3.3110
          0       3.6250
          0       3.9390
     1.0000      4.2530
```

利用 hist 函数自动分为 10 分组，并统计各组频数。由计算结果可知，前 3 组数据和后 3 组数据的频数偏小，故分别将前、后 3 组数据进行合并，这样可得 6 组数据。这 6 组数据所

属的数据组的区间边界值如下：

```
>> a=[];
for k=1:5
    aa=(med(2+k)+med(3+k))/2;
    a=[a,aa];
end
a=[-inf,a,inf]'
```

运行程序，输出如下：

```
a =
      -Inf
    2.2120
    2.5260
    2.8400
    3.1540
    3.4680
       Inf
```

③ 统计经验频数。

②中已经给出经验频数，只需分别合并前、后 3 组的频数。

```
>> f=[f(1)+f(2)+f(3),f(4:7),f(8)+f(9)+f(10)]'
```

运行程序，输出如下：

```
f =
    99
     0
     0
     0
     0
     1
```

④ 计算理论频数。

```
>> PEST=[];
for i=1:6
    pp=normcdf(a(i+1),MU,SIGMA)-normcdf(a(i),MU,SIGMA);
    PEST=[PEST,pp];
end
THEF=n*PEST'
```

运行程序，输出如下：

```
THEF =
   99.4722
    0.5112
```

```
    0.0164
    0.0002
    0.0000
    0.0000
```

⑤ 计算检验统计量的观测值。

```
>> CHI2EST=sum((f-THEF).^2./THEF)
```

运行程序，输出如下：

```
CHI2EST =
   8.6068e+008
```

⑥ 检验决策。

```
>> k=6;
r=2;
alpha=0.1;
df=k-r-1;
REFCR=chi2inv(1-alpha,df);   %拒绝域临界值
p=1-chi2cdf(CHI2EST,df);   %检验的 p 值
if CHI2EST>REFCR
     h=1;
else
     h=0;
end
alpha,h,p,CHI2EST,REFCR
```

运行程序，输出如下：

```
alpha =      0.1000
h =        1
p =        0
CHI2EST =   8.6068e+008
REFCR =      6.2514
```

计算结果表明，在显著性水平 $\alpha=0.10$ 下，h=1 保留拒绝原假设 H_0，即 χ^2 拟合优度检验不认为维尼纶纤度 $X\sim N(1.4042,0.0178^2)$。由最小显著性概率 p=0 表明，当前样本数据下能拒绝原假设，H_0 具有较低的置信水平。

5.5.2 KLJIMOROPOB – CMHPHOB 检验

假设变量 X 的分布函数 $F(x)$ 连续但未知，在给定显著性水平 α 下，要检验假设

$$H_0:\ F(x)=F_0(x);\quad H_1:\ F(x)\neq F_0(x)$$

这个问题可以用 χ^2 拟合优度检验法来检验。

但是，χ^2 拟合优度检验的实质是比较样本频率 $\frac{v_i}{n}$ 与理论频率 $\hat{p}_i=F_0(a_i)-F_0(a_{i-1})$。也

就是说，只是检验了。

$$H_0: \ F(a_i)-F(a_{i-1})=F_0(a_i)-F_0(a_{i-1}), \ i=1,2,\cdots,k$$

其中，a_i 是在连续变量离散化的区间划分过程中得到的。也就是说，只是检验了在区间的分点处 H_0 是否成立而已，这样导致了纳伪风险的增加。于是，人们转而研究更加完善的检验方法。

早在 20 世纪 30 年代初，$\mathrm{K_{LJIMOROPOB}}$ 对分布拟合优度检验问题进行了深入的研究，得到了 $\mathrm{K_{LJIMOROPOB}}$ 定理，进而建立了分布拟合优度检验问题的 $\mathrm{K_{LJIMOROPOB}}$ 检验法和 $\mathrm{C_{MHPHOB}}$ 检验法。

1. $\mathrm{K_{LJIMOROPOB}}$ 检验法

$\mathrm{K_{LJIMOROPOB}}$ 检验法也是比较样本经验函数 $F_n(x)$ 和变量分布函数 $F_0(x)$ 的。但它不是在划分的区间上考虑 $F_n(x)$ 与原假设的分布函数 $F_0(x)$ 之间的偏差，而是在每一点上考虑它们之间的偏差。这就克服了 χ^2 检验法依赖于区间划分的缺点，但其应用范围要窄一些，仅适应于变量的分布函数是连续函数的情形。

根据 $\mathrm{K_{LJIMOROPOB}}$ 定理，当 n 充分大时，样本经验分布函数 $F_n(x)$ 是变量分布函数 $F_0(x)$ 的很好近似，$F_n(x)$ 与 $F_0(x)$ 的偏差一般不应太大。$\mathrm{K_{LJIMOROPOB}}$ 用 $F_n(x)$ 与 $F_0(x)$ 之间的偏差的最大值构造一个统计量

$$D_n=\sup_{-\infty<x<+\infty}\left|F_n(x)-F_0(x)\right|$$

并且得到了下面的定理。

定理 5-2（$\mathrm{K_{LJIMOROPOB}}$ 定理） 设 $X_1,X_2,\cdots,X_n \sim F(x)$ $(n=1,2,\cdots)$，$F(x)$ 为连续的分布函数，在 $F(x)=F_0(x)$（已知）的条件下，有

$$\lim_{x\to\infty} P\left\{D_n<\frac{x}{\sqrt{n}}\right\}=K(x)$$

其中

$$K(x)=\begin{cases}\sum\limits_{k=-\infty}^{+\infty}(-1)^k \mathrm{e}^{-2k^2x^2}, & x>0\\ 0, & x\leqslant 0\end{cases}$$

称为 $\mathrm{K_{LJIMOROPOB}}$ 分布。

根据定理 5-2 检验 H_0：$F(x)=F_0(x)$，若假定 H_0 为真，则当 n 充分大时，检验统计量 $D_n=\sup\limits_{-\infty<x<+\infty}\left|F_n(x)-F_0(x)\right|$ 的值一般应该比较小，若 D_n 的值较大，就应该拒绝 H_0。于是，对给定的显著性水平 α，拒绝域形式为 $D_n\geqslant c$，检验准则为求满足条件 $P\{D_n\geqslant c \mid H_0为真\}\leqslant\alpha$ 的拒绝域临界值 c。

记 $D_n\geqslant D_{n,1-\alpha}$ 为 $\mathrm{K_{LJIMOROPOB}}$ 分布的上侧 α 分位数，即 $P\{D_n\geqslant D_{n,1-\alpha}\}=\alpha$，则 $\mathrm{K_{LJIMOROPOB}}$ 检验法的决策法则是：根据样本数据计算出检验统计量 D_n 的观测值，若

① 当 $D_n\geqslant D_{n,1-\alpha}$ 时，拒绝 H_0，即认为 $F(x)\neq F_0(x)$。

② 当 $D_n<D_{n,1-\alpha}$ 时，接受 H_0，即认为 $F(x)=F_0(x)$。

应用 $\mathrm{K_{LJIMOROPOB}}$ 检验法，原假设 H_0：$F(x)=F_0(x)$ 中的 $F_0(x)$ 的参数应该是已知的。当

参数未知时，对于正态分布，可用参数的大样本估计代替，不过此时的检验是近似的，且显著性水平 α 在 0.1～0.2 之间为宜。

下面概括地给出显著性水平 α 下，用 $K_{\text{LJIMOROPOB}}$ 检验法检验假设



$$H_0:\ F(x)=F_0(x);\ H_1:\ F(x)\neq F_0(x)$$

的步骤。其中，分布函数 $F(x)$ 是连续函数。

① 样本数据排序。将样本数据 $x_1,x_2,\cdots,x_n$（通常 $n\geqslant 50$）按由小到大的次序排列，得到 $x_{(1)}\leqslant x_{(2)}\leqslant\cdots\leqslant x_{(n)}$。

② 求出经验分布函数。

$$F_n(x)=\begin{cases}0, & x<x_{(1)}\\ \dfrac{1}{n}\sum\limits_{i=1}^{k}v_i, & x_{(k)}\leqslant x<x_{(k+1)}(k=1,2,\cdots,n-1)\\ 1, & x\geqslant x_{(n)}\end{cases}$$

其中，v_i 为样本数据 $x\in[x_{(i)},x_{(i+1)})$ 的频数，且 $\sum v_i=n$。

③ 计算检验统计量 D_n 的值。

$D_n=\sup\limits_{-\infty<x<+\infty}|F_n(x)-F_0(x)|=\max\limits_{\forall i}\left\{|F_n(x_{(i)})-F_0(x_{(i)})|,|F_n(x_{(i+1)})-F_0(x_{(i)})|\right\}$。其中，规定 $F_n(x_{(n+1)})=1$。

④ 求 $K_{\text{LJIMOROPOB}}$ 分布的上侧 α 分位数 $D_{n,1-\alpha}$，当 $n>100$ 时，常用 $D_{n,1-\alpha}$ 近似公式如下：

$$D_{n,0.80}\approx 1.07/\sqrt{n},\quad D_{n,0.90}\approx 1.23/\sqrt{n},\quad D_{n,0.95}\approx 1.36/\sqrt{n},\quad D_{n,0.99}\approx 1.63/\sqrt{n}$$

⑤ 检验决策

若 $D_n\geqslant D_{n,1-\alpha}$，则拒绝 H_0，认为样本数据不是来自理论分布 $F_0(x)$ 的。

若 $D_n<D_{n,1-\alpha}$，则接受 H_0，认为样本数据是来自理论分布 $F_0(x)$ 的。

2. C_{MHPHOB} 检验法

C_{MHPHOB} 检验法是对 $K_{\text{LJIMOROPOB}}$ 检验法的一种推广。

设 $X_1,X_2,\cdots,X_n\sim F(x)$，$Y_1,Y_2,\cdots,Y_m\sim G(x)$ $(n,m=1,2,\cdots)$，$F(x)$ 和 $G(x)$ 均为连续的分布函数，$-\infty<x<+\infty$，在显著性水平 α 下，检验假设

$$H_0:\ F(x)=G(x);\ H_1:\ F(x)\neq G(x)$$

用 $F(x)$ 和 $G_m(x)$ 分别表示两样本的经验分布函数，用它们构造检验统计量

$$D_{nm}=\sup_{-\infty<x<+\infty}|F_n(x)-G_m(x)|$$

C_{MHPHOB} 证明了下面的定理。

定理 5-3（$K_{\text{LJIMOROPOB}}-C_{\text{MHPHOB}}$ 定理） 当 H_0 为真且样本容量 n 和 m 分别趋向于 ∞ 时，有

$$\lim_{n,m\to\infty}P\left\{\sqrt{\frac{nm}{n+m}}D_{nm}<x\right\}=K(x)$$

其中，$K(x)$ 是 $K_{\text{LJIMOROPOB}}$ 分布函数。

根据定理 5-3，可得检验 H_0：$F(x)=G(x)$ 的 C_{MHPHOB} 检验法则（近似）：

① 若 $D_{nm} \geqslant D_{nn,1-\alpha}$，则拒绝 H_0，认为 $F(x) \neq G(x)$。

② 若 $D_{nm} < D_{nn,1-\alpha}$，则接受 H_0，认为 $F(x)=G(x)$。

在应用中，确定 $K_{\text{LJIMOROPOB}}$ 分布的分位数 $D_{nn,1-\alpha}$ 时，用 $N=\left[\dfrac{nm}{n+m}\right]$ 代替前述分位数近似公式中的 n，而计算 D_{nm} 的观测值用下面的公式：

$$D_{nm}=\max_{\forall i}\left|F_n(x_{(i)})-G_m(x_{(i)})\right|$$

其中，x_i 为划分变量值域的第 i 个小区间的组中值。

MATLAB 将这两种检验方法统称为 $K_{\text{LJIMOROPOB}}-C_{\text{MHPHOB}}$（英文书写为 Kolmogorov–Smirnov）检验，并提供了两个检验函数 kstest 和 kstest2。

（1）kstest 函数

kstest 函数用于大样本情形下连续变量分布形态的拟合优度检验。

其调用格式如下：

```
[h,p,stats,cv]=kstest(x,cdf,alpha,tail)
```

其中，输入参数 x 为样本数据向量，cdf 为检验的原假设所指定的分布形式（具体引用为变量的累积分布函数，默认时 cdf=[]，表示拟合标准正态分布），alpha 为检验的显著性水平（默认时为 0.05），tail 为备择假设类型的标示值。输出参数 h 为检验决策，p 为拒绝原假设的最小显著性概率，stats 为检验统计量的值，cv 为拒绝域的临界值。

（2）kstest2 函数

kstest2 函数用于大样本情形下两个连续变量分布一致性的检验。

其调用格式如下：

```
[h, p, stats]=kstest2(x, y, alpha, tail)
```

检验的原假设是两个变量服从相同的分布。输入参数 x 和 y 分别为两个样本的数据向量，其他输入、输出参数的意义同 kstest 函数。

【例 5-17】 在显著性水平 $\alpha=0.10$ 下，用 $K_{\text{LJIMOROPOB}}-C_{\text{MHPHOB}}$ 检验法对例 5-15 中的维尼纶纤度数据进行正态性检验。

其实现的 MATLAB 程序代码如下：

```
>>clear all;
load data
[MU,SIGMA]=normfit(data)
x=(data-MU)/SIGMA;
[h,p,stats,cv]=kstest(x,[],0.1,0)
```

运行程序，输出如下：

```
MU =      1.4338
SIGMA =       0.3043
```



```
h =        1
p =   6.3167e-014
stats =      0.3897
cv =       0.1207
```

结果表明，接受原假设，即认为维尼纶纤度服从均值为 1.4338、标准差为 0.3043 的正态分布。

5.5.3 正态性检验

检验变量是否服从正态分布是统计应用中最常见的，也是最重要的问题。此类问题当然可以用 $K_{LJIMOROPOB}-C_{MHPHOB}$ 检验法进行，但是，由于受样本容量因素的影响，有时检验效果可能不理想。因此，人们发现了一些专门的正态性检验方法，其检验效果一般比通用方法好。这里介绍 3 种常用的正态性检验方法。

1. 正态概率纸检验法

正态概率纸是一种现场统计常用的判断变量正态性的简单工具，使用它可以很快地判断变量是否服从正态分布，还能够粗略地估计出分布的数字特征。

首先介绍正态概率纸的构造原理。

设变量 X 的分布函数为 $F(x)$，需要检验

$$H_0:\ X \sim N(\mu,\sigma^2),\ -\infty<\mu<+\infty, \sigma^2>0$$

在原假设 H_0 成立时，$\dfrac{X-\mu}{\sigma}=U \sim N(0,1)$，而且 $F(x)$ 可用标准正态分布 $N(0,1)$ 的分布函数 $\Phi(x)$ 来表示

$$F(x)=\Phi\left(\frac{X-\mu}{\sigma}\right)=\Phi(u)$$

其中

$$u=\frac{1}{\sigma}(x-\mu)$$

在 xOu 直角坐标平面上，假定横轴（x 轴）与纵轴（u 轴）的单位长度相等，函数 $u=\dfrac{1}{\sigma}(x-\mu)$ 的图像是一条直线，经过点 $(\mu,0)$，斜率为 $1/\sigma$。

为使这条直线能够直观地解释变量的取值 x 与 $P\{X\leqslant x\}$ 之间的关系，进行如下坐标刻度更新：在直角坐标系 xOu 中，保持横轴上 x 的刻度不变，而把纵轴上 u 的刻度更新为 $y=100\Phi(u)$，并规定 $100\Phi(-\infty)=0$，$100\Phi(+\infty)=100$。这样就将直角坐标系 xOu 更新为直角坐标系 xOy。由于 y 轴上的刻度 0 与 100 分别对应 u 轴上的 $-\infty$ 和 $+\infty$，因此 y 轴上无法标示出 0 与 100，一般 y 轴上的刻度标示限于 0.01～99.99 之间。称以直角坐标系 xOy 为刻度体系的坐标纸为正态概率纸。

根据正态概率纸的构造原理可知，xOu 直角坐标系中 x 与 u 的关系，在 xOy 直角坐标系中就成为 x 与 $y=100P\{X\leqslant x\}$ $(=100F(x)=100\Phi(u))$ 的关系；反之亦然。特别对于正态概率纸上的一条直线，若该直线能表示为 $u=\dfrac{1}{\sigma}(x-\mu)$，则 $100F(x)$ 与 x 的关系为

$$100F(x)=100\Phi(u)=100\Phi\left(\frac{x-\mu}{\sigma}\right)$$

即

$$F(x)=\Phi\left(\frac{x-\mu}{\sigma}\right)$$

也就是说，$F(x)$ 是一个正态分布的分布函数。

这表明，正态概率纸上斜率存在且大于零的全体直线所组成的集合与全体正态分布函数所组成的正态分布族之间存在一一对应关系。

2. Lilliefors 检验

Lilliefors 检验法是对 $K_{LJIMOROPOB}$ 检验法的一种改进。

设 $X_1,X_2,\cdots,X_n \sim X$，$X$ 的分布未知。需要检验

$$H_0:\ X\sim N(\mu,\sigma^2),\ -\infty<\mu<+\infty,\sigma^2>0$$

令 $\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$，$S=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2}$，$Z_i=\frac{X_i-\bar{X}}{S}$ $(i=1,2,\cdots,n)$，则当 H_0 为真时，标准化样本 $Z_1,Z_2,\cdots,Z_n \text{i.i.d.}\sim N(0,1)$，于是 $K_{LJIMOROPOB}$ 统计量可修正为

$$D_n=\sup_{-\infty<x<+\infty}\left|S_n(x)-\Phi(x)\right|$$

其中，$S_n(x)$ 是标准化样本的经验分布函数。这就是 Lilliefors 检验的检验统计量。

其他如检验法则、检验步骤等与 $K_{LJIMOROPOB}$ 检验法类似，这里不再介绍。

由 Lilliefors 检验的检验统计量的构造特点可知，该方法与 $K_{LJIMOROPOB}$ 检验法最大的不同之处是检验不需要已知分布参数，样本的标准化避免了在正态拟合优度检验之前对分布参数的估计，因此该方法可在小样本条件下使用。

MATLAB 提供了 Lilliefors 检验法的检验函数 lillietest。

其调用格式如下：

```
[h,p,stats,cv]=lillietest(x, alpha, tail)
```

其输入、输出参数的意义同 kstest 函数。

3. Jarque–Bera 检验

Jarque–Bera 检验是一种常用的、基于峰度与偏度联合检验的正态性检验方法。

设 $X_1,X_2,\cdots,X_n \sim X$，$X$ 的分布未知。需要检验

$$H_0:\ X\sim N(\mu,\sigma^2),\ -\infty<\mu<+\infty,\sigma^2>0$$

令 $B_k=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^k$，Jarque 和 Bera 由样本峰度 $KU=\frac{B_4}{B_2^2}$ 和样本偏度 $SK=\frac{B_3}{B_2^{3/2}}$ 定义了如下的检验统计量：

$$J=\frac{n}{6}\left[SK^2+\frac{(KU-3)^2}{4}\right]$$

并证明了在 H_0 为真的条件下，J 渐近地服从自由度为 2 的 χ^2 分布。

由于正态分布的样本峰度 $KU=3$，样本偏度 $SK=0$，因此检验统计量 J 的观测值越大越对 H_0 不利。于是，对于给定的显著性水平 α，检验准则为 $P\{J>\chi^2_{1-\alpha}(2)\}\leqslant\alpha$。当检验统计量的实测值 $J>\chi^2_{1-\alpha}(2)$ 时，则在显著性水平 α 下拒绝原假设 H_0，否则保留 H_0。



由于检验依据是渐近分布，因此该方法应在大样本条件下使用。

MATLAB 提供了 Jarque–Bera 检验法的检验函数 jbtest。

其调用格式如下：

```
[h,p,stats,cv]=jbtest(x,alpha,tail)
```

其输入、输出参数的意义同 kstest 函数。

【例 5-18】 某工厂生产一种白炽灯，其流明为随机变量 ξ，假设 ξ 满足正态分布 $N(\mu,\sigma^2)$，现从产品中随机抽取 120 个样本，其指标（流明数）如下，试检验正态分布的假设是否正确。

```
216,203,197,208,206,209,206,208,202,203,206,213,218,207,208,202,194,203,213,211
193,213,208,208,204,206,204,206,208,209,213,203,206,207,196,201,208,207,213,208
210,208,211,211,214,220,211,203,216,224,211,209,218,214,219,211,208,221,211,218
218,190,219,211,208,199,214,207,207,214,206,217,214,201,212,213,211,212,216,206
210,216,204,221,208,209,214,214,199,204,211,201,216,211,209,208,209,202,211,207
202,205,206,216,206,213,206,207,200,198,200,202,203,208,216,206,222,213,209,219
```

其实现的 MATLAB 程序代码如下：

```
>> clear all;
x=[216,203,197,208,206,209,206,208,202,203,206,213,218,207,208,202,194,203,213,211,...
   193,213,208,208,204,206,204,206,208,209,213,203,206,207,196,201,208,207,213,208,...
   210,208,211,211,214,220,211,203,216,224,211,209,218,214,219,211,208,221,211,218,...
   218,190,219,211,208,199,214,207,207,214,206,217,214,201,212,213,211,212,216,206,...
   210,216,204,221,208,209,214,214,199,204,211,201,216,211,209,208,209,202,211,207,...
   202,205,206,216,206,213,206,207,200,198,200,202,203,208,216,206,222,213,209,219];
[h,p]=jbtest(x,0.05)
```

运行程序，输出如下：

```
h =      0
p =    0.5000
```

确定了该数据为正态分布数据，则可以直接用前面介绍的正态分布拟合函数 normfit 求出该分布的均值、方差及其置信区间。

```
>> [mu1,sig1,mu_ci,sig_ci]=normfit(x,0.05);
mu=[mu1,mu_ci']
```

运行程序，输出如下：

```
mu =
   208.8167   207.6737   209.9596
```

```
>> sig=[sig1,sig_ci']
```

运行程序，输出如下：

```
sig =
     6.3232     5.6118     7.2428
```

【例 5-19】 在显著性水平 $\alpha=0.10$ 下，分别用正态概率纸检验法、Lilliefors 检验法和 Jarque–Bera 检验法对例 5-15 中的维尼纶纤度数据进行正态性检验。

其实现的 MATLAB 程序代码如下：

```
>> clear all;
load data
```

（1）正态概率纸检验法

```
>> normplot(data)
```

运行程序，效果如图 5-1 所示。

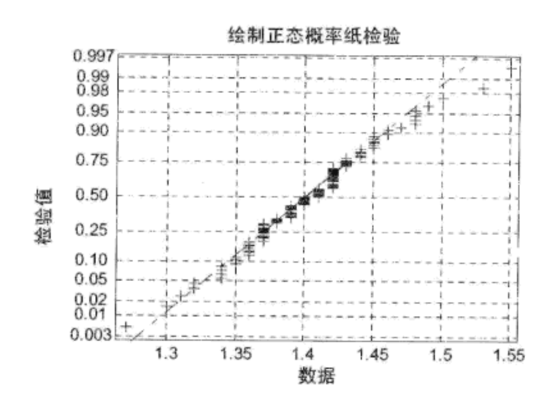


图 5-1 正态概率纸检验图

（2）Lilliefors 检验法

```
> [L1,L2]=lillietest(data,0.1)
```

运行程序，输出如下：

```
L1 =      1
L2 =    0.0659
```

（3）Jarque-Bera 检验法

```
>> [J1,J2]=jbtest(data,0.1)
```

运行程序，输出如下：

```
J1 =      0
```



```
J2 =     0.2038
```

从图 5-1 中可以看出，100 个样本数据的$(x_i,100F_n(x_i))$点列在一直线附近，故可认为维尼纶纤度数据来自正态分布。从图 5-1 中可以粗略地估计出维尼纶纤度的均值约为 1.4，标准差约为 1.45 − 1.4 = 0.05。



Jarque-Bera 检验法的结论是接受维尼纶纤度服从正态分布的假设。

值得注意的是，Lilliefors 检验法得到的结论是拒绝维尼纶纤度服从正态分布的假设，这是由于样本数据的标准化变换，使得该方法对异常数据（极端数据）反应敏感。其实，若注意到第 99 个数据$x_{99}=1.55$是 data 数据集中的最大值，从正态概率纸检验的图形中可以看出这个最大值过于偏离直线$y=\frac{1}{\sigma}(x-\mu)$，所以$x_{99}$是一个异常数据。若从 data 数据集中删除这个数据，重新进行检验，如下所示。

```
>> data(99)=[];
[h,p]=lillietest(data,0.1)
```

运行程序，输出如下：

```
h =      0
p =     0.1662
```

结果表明，剩余的 99 个维尼纶纤度数据是来自正态分布的，与另外两种检验方法的结论一致。

5.5.4 符号检验法

如果两个随机变量 x 和 y 具有相同的概率分布，对它们各进行 n 次测量，得到两组样本值：$x_1,x_2,\cdots,x_n$ 和 $y_1,y_2,\cdots,y_n$。

如果两组样本值之间不存在系统误差，那么出现$x_i>y_i$与出现$x_i<y_i$的机会是相等的，概率各为 1/2。

统计工具箱提供了零中值分布的符号检验函数 signtest。

其调用格式如下：

```
p = signtest(x)
p = signtest(x,m)
p = signtest(x,y)
[p,h] = signtest(...)
[p,h] = signtest(...,'alpha',alpha)
[p,h] = signtest(...,'method',method)
[p,h,stats] = signtest(...)
```

其中，x，y 是分析的样本；m 是设定的中值；alpha 是显著性水平；method 为实现检验的方法。

下面通过举例来说明符号检验法的应用。

【例 5-20】 零中值分布的符号检验。

其实现的 MATLAB 程序代码如下：

```
>> clear all;
N=1024;            %样本点数
x1=randn(1,N);     %正态分布
%假设检验
alpha0=0.05;
[p1,h1]=signtest(x1,alpha0)
%Weibull 分布
x2=wblrnd(1,2,N,1);
%实现的假设检验
[p2,h2]=signtest(x2,alpha0)
%两个样本的差
```

运行程序，输出如下：

```
p1 =   0.0749
h1 =       0
```

即接受 $x1$ 来自于零中值的分布。

```
p2 =   1.6065e-222
h2 =       1
```

即不接受 $x2$ 来自于零中值的分布。

```
p3 =   4.3185e-087
h3 =       1
```

即不接受 $x1$ 和 $x2$ 的差服从零中值的分布。

5.5.5 秩和检验法

如果两个总体 A,B 具有相同的概率分布，分别从两个总体 A,B 中抽取大小为 n_1 和 n_2 的样本进行测定，得到两组测定值：$x_1,x_2,\cdots,x_{n_1}$ 和 $y_1,y_2,\cdots,y_{n_2}$。

将两组测定值混合起来按由小到大的顺序排列，每个测定值在序列中排列的次序，称为该测定值的秩。一组样本测定值中各测定值的秩的总和，称为该组测定值的秩和。当两组测定值中某个值相等时，其秩等于相应两个测定值秩的平均值。如果两个总体具有相同的概率分布，那么在混合排列的序列中第 i 个序次为测定值 x_i 或 y_i 的概率是相同的。

统计工具箱提供了两个样本服从同中值分布的秩和检验函数 ranksum。

其调用格式如下：

```
p = ranksum(x,y)
[p,h] = ranksum(x,y)
[p,h] = ranksum(x,y,'alpha',alpha)
[p,h] = ranksum(...,'method',method)
[p,h,stats] = ranksum(...)
```

其中，x，y 是分析的样本；alpha 是显著性水平。p 为返回概率；h 为返回假设检验结果。

下面通过举例来说明秩和检验法的应用。



【例 5-21】 两个样本是否服从同中值分布的秩和检验。

其实现的 MATLAB 程序代码如下：

```
>> clear all;
N1=100; N2=150;   %样本点数
alpha0=0.05;
x=unifrnd(0,1,N1,1);
y=unifrnd(0.25,1.25,N2,1);
%假设检验
[p1,h1]=ranksum(x,y,alpha0)
[p2,h2]=ranksum(x,x,alpha0)
[p3,h3]=ranksum(y,y,alpha0)
```

运行程序，输出如下：

```
p1 =   2.9883e-010
h1 =        1
```

即不接受 x 和 y 来自于相同中值的分布，这与实际是一致的；x 和 y 来自于同一分布，但参数偏移 0.25。

```
p2 =        1
h2 =        0
p3 =        1
h3 =        0
```

即接受 x 和 x，y 和 y 来自于相同中值的分布，这与实际是一致的。

第 6 章　方差分析及曲线拟合

方差分析是重要的、应用广泛的实验数据统计分析方法，其实质是检验多个变量均值的一致性。由于检验的统计推断是通过讲解实验数据的变异性以及变异的来源作出的，而统计分析刻画数据变异性的基本统计量是样本方差，因此，习惯上称这种多变量均值一致性的假设检验为方差分析。下面对方差分析及曲线拟合作介绍。

6.1　方差分析的相关概念

6.1.1 基本概念

在实际中，常常要通过实验来了解各种因素对产品的性能、产量等的影响，这些性能、产量等指标统称为实验指标，而称影响实验指标的条件、原因等为因素或因子，称因素所处的不同状态为水平。各因素对实验指标的影响一般是不同的，就是一个因素的不同的水平对实验指标的影响往往也是不同的。方差分析就是通过对实验数据进行分析，检验方差相同的各正态总体的均值是否相等，以判断各因素对实验指标的影响是否显著。方差分析按影响实验指标的因素的个数分为单因素方差分析、双因素方差分析和多因素方差分析。下面将对它们展开介绍。

在实验研究中，所获得的实验结果（数据）总是有差异的，即使在同一条件下重复进行实验，所得实验数据也不完全一样，引起实验数据产生差异的因素很多，这些因素对实验数据的影响程度也是不同的，有主有次，有大有小。通常，由于因素变化所引起的数据差异称为条件误差，它决定了实验结果的准确度。在实验过程中，由于一系列有关因素的细小、随机（偶然）的波动而形成的具有相互抵消性的误差称为随机误差，它决定了实验结果的精密度。

6.1.2 方差分析的必要性

在前面介绍中，已经讲解了两个样本均值相等的假设实验问题。在生产实践中，经常遇到多个样本均值是否相等的问题。

【例 6-1】 在以淀粉为原料生产葡萄糖的过程中，残留了许多糖蜜，可作为生产酱色的原料。在生产酱色之前应尽可能彻底除杂，以保证酱色质量，为此对除杂方法进行选择。在实验中选用 5 种不同的除杂方法，每种方法做 4 次实验，即重复 4 次，结果见表 6-1 所示。

表 6-1　不同除杂的除杂量　　　　（单位：g/kg）

除杂方法 A_i	除杂量 x_{ij}				平均量 $\bar{x}_i$
A_1	25.6	22.2	28.0	29.8	26.4
A_2	24.4	30.0	29.0	27.5	27.7
A_3	25.0	27.7	23.0	32.2	27.0
A_4	28.8	28.0	31.5	25.9	28.6
A_5	20.6	21.2	22.0	21.2	21.3



本实验的目的是判断不同的除杂方法对除杂量是否有显著影响，以便确定最佳除杂方法。从表 6-1 可见，各次实验结果是参差不齐的。可以认为，同一除杂方法重复实验得到的 4 个数据的差异是由随机误差造成的，而随机误差常常是服从正态分布的，这时除杂量应该有一个理论上的均值。而对不同的除杂方法，除杂量应该有一个不同的均值。这种均值之间的差异是由除杂方法的不同造成的。于是可以认为，5 种除杂方法下所得数据是来自均值不同的 5 个正态总体，且由于实验中其他条件相对稳定，因而可以认为每个总体的方差是相同的，即 5 个总体具有方差齐性。这样，判断除杂方法对除杂效果是否有显著影响的问题，就转化为检验 5 个具有相同方差的正态总体的均值是否相同的问题了，即检验假设

$$H_0:\ \mu_1=\mu_2=\mu_3=\mu_4=\mu_5$$

在上述这种情况下，第 5 章介绍的方法不再适用。这是因为：1）倘若是 10 个样本，需要检验 H_0：$\mu_1=\mu_2,\mu_3=\mu_4,\cdots,\mu_9=\mu_{10}$，共需检验 $\dfrac{k(k-1)}{2}=45$ 个假设，这样的程序非常烦琐。2）样本进行两两比较时，只能由 $2(n-1)$ 个自由度估计样本均值、标准误差，而不能由 $10(n-1)$ 个自由度一起估计，精度不够高。3）两两检验会随着样本个数的增加而大大增加错误的可能性。比如，在两两比较中 α 取 0.05，45 次比较的结论都正确的概率为 0.95^{45}，至少做出一次错误的结论的概率为 $1-0.95^{45}=0.9006$，这时的检验结果已经很不可靠。对于这种多个总体样本均值的假设检验，需采用方差分析方法。

6.1.3 方差分析的基本思想

方差分析的实质就是检验多个正态总体的均值是否相等。那么如何检验呢？从表 6-1 可见，20 个数据是参差不齐的，数据波动的可能原因来自两个方面：一是由于因素的水平不同，即除杂方法不同造成的。事实上，5 种除杂方法下的数据平均值 $\bar{x}_i$ 之间确实有差异。二是来自偶然误差，从表中数据可见，每一种除杂方法下的 4 个数据虽然是相同条件下的实验结果，但仍然存在差异，这是由实验中存在的偶然因素（例如，环境、原材料成分、测试技术等微小而随机的变化）引起的。这里，把由因素的水平变化引起的实验数据波动称为条件误差；把随机因素引起的实验数据波动称为随机误差或实验误差。方差分析就是把实验数据的总波动分解为两个部分，一部分反映由条件误差引起的波动，另一部分反映由实验误差引起的波动。亦即把数据的总偏差平方和 S_T 分解为反映必然性的各个因素的偏差平方和 $S_A,S_B,\cdots$ 与反映偶然性的偏差平方和 S_e，并计算它们的平均偏差平方和。再将两者进行比较，借助 F 检验法，检验假设 H_0：$\mu_1=\mu_2=\cdots$，从而确定因素对实验结果的影响是否显著。也就是说，方差分析所分析的并非方差，而是研究数据间的变异来源是条件误差还是随

机误差。

为方便说明方差分析的基本思想与方法，下面考查一个简单的、易于理解的例子。

【例 6-2】 一位英语教师想检查 3 种不同的教学方法的效果，为此随机选取 24 名学生并把他们分成 3 组，相应地用 3 种方法教学。一段时间后，这位教师对这 24 名学生进行统考，统考成绩见表 6-2。试问在显著性水平$\alpha = 0.05$下，这 3 种教学方法有无显著性差异？

表 6-2 英语成绩表

方法	学习成绩								
A_1	73	66	89	82	43	80	63		
A_2	88	78	91	76	85	84	80	96	
A_3	68	79	71	71	87	68	59	76	80

表 6-2 中，A_1，A_2，A_3是这位英语教师采用的不同教学方法，各有其侧重点。目的是判断不同教学方法对英语学习成绩是否有显著影响。若有影响，哪一种教学方法好？

容易理解，在不同的教学方法下，学生的英语成绩可能是不同的；在同一种方法下，不同学生的英语成绩也可能是不同的。也就是说，实验数据是有差异的，而差异可能是由因素的不同处理（3 种不同的教学方法）引起的，这种差异称为实验数据的条件误差；也可能是由随机因素（不可控制或不可预知的因素，如考试时的环境、时间对学生的影响）引起的，这种差异称为实验数据的随机误差或实验误差。方差分析的主要任务就是推断在因素的不同处理下，响应变量的均值（3 种不同教学方法下学生的英语平均成绩）是否一致，而进行推断的基本思想就是分析实验数据的差异来源。在后面的讲解中可以看到，其中关键性的思想方法是考查实验数据的偏差平方和，并设想将数据总的偏差平方和按照产生的原因分解成“总偏差平方和=条件误差平方和+随机误差平方和”，然后进一步比较这两种偏差平方和的大小，按照一定的统计假设检验的规则确定总的差异（总偏差平方和）究竟是由条件误差（因素的不同处理引起的偏差平方和），还是随机误差（随机因素引起的偏差平方和）决定的。如果实验数据的差异是由条件误差决定的，则说明在因素的不同处理下响应变量的均值是不同的；如果差异不是由条件误差决定的，则在因素的不同处理下响应变量的均值应当是一致的。

6.2 单因素方差分析

6.2.1 单因素统计模型及检验方法

1. 统计模型

例 6-2 中所考查的因素只有一个，称其为单因素试验。通常在单因素试验中，设因素 A 有 r 个水平 $A_1, A_2, \cdots, A_r$（即试验中有 r 个处理），在每一水平下考查的指标可以看成一个变量。现有 r 个水平，故有 r 个变量。为简化起见，需要给出若干假定，把所要回答的问题归结为一个统计问题，然后设法解决它。假定：

1）每一变量均服从正态分析。

2）每一变量的方差相同。

3）从 r 个变量抽取的样本相互独立。

要比较各个变量的均值是否一致，设第 i 个变量的均值为 μ_i，那么就要检验如下假设：

$$H_0: \ \mu_1 = \mu_2 = \cdots = \mu_r$$



其备择假设为

$$H_1: \ \mu_1, \mu_2, \cdots, \mu_r$$

其中，$\mu_1, \mu_2, \cdots, \mu_r$ 不全相同，通常 H_1 可以省略不写。

当 H_0 为真时，称因素 A 的各水平间无显著差异，简称因素 A 不显著（此时在例 6-2 中，得出不同的教学方法对英语学习成绩没有显著影响）；当 H_0 不为真时，各 $\mu_i \ (i = 1, 2, \cdots, r)$ 不全相同，这时称因素 A 的各水平间有显著差异，简称因素 A 显著。

用于检验假设 H_0 的统计方法称为方差分析法，其实质上是检验若干个具有相同方差的正态变量的均值是否相等的一种统计方法。在所考虑的因素仅有一个的场合，称为单因素方差分析。

为检验假设 H_0，需要对每一变量抽取样本。这些样本可以通过试验或某种观察获得。各样本间还是相互独立的。为方便起见，本章对样本及其观察值都用符号 y 加下标表示，其含义可从上下文理解。设第 i 个变量对应容量为 m_i 的样本 $y_{i1}, y_{i2}, \cdots, y_{im_i} \ (i = 1, 2, \cdots, r)$。

在 A_i 水平下获得的 y_{ij} 与 μ_i 不会总是一致的，如例 6-2 中教学方法 A_1 下学生的成绩也不完全相同。记为

$$\varepsilon_{ij} = y_{ij} - \mu_i$$

称 ε_{ij} 为随机误差，从而有

$$y_{ij} = \varepsilon_{ij} + \mu_i$$

称上式为 y_{ij} 的数据结构式，即均值为 μ_i 的变量观察值 y_{ij} 可看成是由其均值 μ_i 与随机误差 ε_{ij} 叠加而产生的。假定 A_i 的指标 y_{ij} 服从 $N(\mu_i, \sigma^2)$ 分布，则有 $\varepsilon_{ij} \sim N(0, \sigma^2)$。

综上，有单因素方差分析的统计模型：假定

$$\left.\begin{array}{c} y_{ij} = \varepsilon_{ij} + \mu_i, \\ \varepsilon_{ij} \sim N(0, \sigma^2)\text{且相互独立}, \end{array} \quad i = 1, 2, \cdots, r; j = 1, 2, \cdots, m_i \right\} \tag{6-1}$$

检验假设 H_0： $\mu_1 = \mu_2 = \cdots = \mu_r$。

为了能更仔细地描述数据，常在方差分析模型中引入一般平均与效应的概念。称诸 μ_i 为加权平均。称

$$\mu = \frac{1}{n} \sum_{i=1}^{r} m_i \mu_i$$

为一般平均，其中 $n = \sum_{i=1}^{r} m_i$，称

$$a_i = \mu_i - \mu, i = 1,2,\cdots,r$$

为因素 A 的第 i 水平的主效应，也简称为 A_i 的效应。容易看出，效应间有如下关系式：

$$\sum_{i=1}^{r} m_i a_i = 0$$

在上述记号下，有

$$\mu_i = a_i + \mu$$

这表明第 i 个总体的均值是一般平均与其效应的叠加。此时，单因素方差分析的统计模型可改写成

$$\begin{cases} y_{ij} = \mu + \varepsilon_{ij} + a_i, \\ \sum_{i=1}^{r} m_i a_i = 0, \qquad i = 1,2,\cdots,r; j = 1,2,\cdots,m_i \\ \varepsilon_{ij} \sim N(0,\sigma^2) \text{且相互独立,} \end{cases} \tag{6-2}$$

它由数据结构式、关于效应的约束条件及关于误差的假定 3 部分组成。在上述模型下，所要检验的假设可改写成

$$H_0: \ a_1 = a_2 = \cdots = a_r = 0$$

2. 检验方法

为了使差异的大小能定量地表示出来，先引入如下若干记录。

把 A_i 水平下的试验数据和记为 $y_{i\cdot} = \sum_{j=1}^{m_i} y_{ij}$，其平均值记为 $\overline{y}_{i\cdot} = \frac{1}{m_i} y_{i\cdot}$，由 y_{ij} 的数据结构式可知，$\overline{y}_{i\cdot}$ 具有如下结构式：

$$\overline{y}_{i\cdot} = \mu_i + \overline{\varepsilon}_{i\cdot}$$

其中，$\overline{\varepsilon}_{i\cdot} = \frac{1}{m_i}\sum_{j=1}^{m_i} \varepsilon_{ij}$。

把所有数据之和记为 $y_{\cdot\cdot} = \sum_{i=1}^{r}\sum_{j=1}^{m_i} y_{ij}$，其平均值记为 $\overline{y} = \frac{y_{\cdot\cdot}}{n}$，$\overline{y}$ 具有如下结构式：

$$\overline{y} = \mu + \overline{\varepsilon}$$

其中，$\overline{\varepsilon} = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{m_i} \varepsilon_{ij}$。由于

$$y_{ij} - \overline{y} = (y_{ij} - \overline{y}_{i\cdot}) + (\overline{y}_{i\cdot} - \overline{y})$$

其中，$y_{ij} - \overline{y}_{i\cdot}$ 称为组内偏差，仅反映随机误差：

$$y_{ij} - \overline{y}_{i\cdot} = (\mu_i + \varepsilon_{ij}) - (\mu_i + \overline{\varepsilon}_{i\cdot}) = \varepsilon_{ij} - \overline{\varepsilon}_{i\cdot}$$

而 $\overline{y}_{i\cdot} - \overline{y}$ 称为组间偏差，除了反映随机误差外，还反映了第 i 个水平效应：

$$\overline{y}_{i\cdot} - \overline{y} = (\mu_i + \overline{\varepsilon}_{i\cdot}) - (\mu + \overline{\varepsilon}) = a_i + \overline{\varepsilon}_{i\cdot} - \overline{\varepsilon}$$



各 y_{ij} 间总的差异大小可用总偏差平方和 SST 表示：

$$SST = \sum_{i=1}^{r}\sum_{j=1}^{m_i}(y_{ij} - \overline{y})^2$$

由随机误差引起的数据间的差异可以用组内偏差平方和表示。由于组内偏差仅反映随机误差，故也把组内偏差平方和称为误差偏差平方和，记为 SSE：

$$SSE = \sum_{i=1}^{r}\sum_{j=1}^{m_i}(y_{ij} - \overline{y}_{i\cdot})^2$$

由于组间偏差除了反映随机误差外，还反映了效应间的差异，故由效应不同引起的数据差异可用组间偏差平方和表示，也称为因素 A 的偏差平方和，记为 SSA：

$$SSA = \sum_{i=1}^{r} m_i(\overline{y}_{i\cdot} - \overline{y})^2$$

这里，每一项乘上 m_i 是因为第 i 水平有 m_i 个实验数据。

定理 6-1（平方和分解定理 1） $SST=SSA+SSE$。

事实上

$$\begin{aligned}SST &= \sum_{i=1}^{r}\sum_{j=1}^{m_i}(y_{ij} - \overline{y})^2 = \sum_{i=1}^{r}\sum_{j=1}^{m_i}(y_{ij} - \overline{y}_{i\cdot} + \overline{y}_{i\cdot} - \overline{y})^2 \\ &= \sum_{i=1}^{r}\sum_{j=1}^{m_i}(y_{ij} - \overline{y}_{i\cdot})^2 + \sum_{i=1}^{r}\sum_{j=1}^{m_i}(\overline{y}_{i\cdot} - \overline{y})^2 + 2\sum_{i=1}^{r}\sum_{j=1}^{m_i}(y_{ij} - \overline{y}_{i\cdot})(\overline{y}_{i\cdot} - \overline{y}) \\ &= SSE + SSA\end{aligned}$$

由于 $\sum_{j=1}^{m_i}(y_{ij} - \overline{y}_{i\cdot}) = 0$，故上述第三项为 0。

由模型（6-2）（即式（6-2））可知各 ε_{ij} 相互独立，且 $\varepsilon_{ij} \sim N(0,\sigma^2)$ $(i = 1,2,\cdots,r;$ $j = 1,2,\cdots,m_i)$，故

$$\overline{\varepsilon}_{i\cdot} \sim N\left(0,\frac{\sigma^2}{m_i}\right),\ i = 1,2,\cdots,r$$

$$\overline{\varepsilon} \sim N\left(0,\frac{\sigma^2}{N}\right)$$

由于

$$\frac{1}{\sigma^2}\sum_{j=1}^{m_i}(y_{ij} - \overline{y}_{i\cdot})^2 = \frac{1}{\sigma^2}\sum_{j=1}^{m_i}(\varepsilon_{ij} - \overline{\varepsilon}_{i\cdot})^2 \sim \chi^2(m_i - 1)$$

又由 χ^2 分布的可加性可知

$$\frac{SSE}{\sigma^2} = \sum_{i=1}^{r}\left[\frac{1}{\sigma^2}\sum_{j=1}^{m_i}(y_{ij} - \overline{y}_{i\cdot})^2\right] \sim \chi^2\left(\sum_{i=1}^{r}(m_i - 1)\right) = \chi^2(n - r)$$

由 χ^2 分布的性质知

$$E\left(\frac{SSE}{\sigma^2}\right)=n-r$$

即

$$E(SSE)=(n-r)\sigma^2$$

由于

$$SSA=\sum_{i=1}^{r}m_i(\overline{y}_{i\cdot}-\overline{y})^2=\sum_{i=1}^{r}m_i(a_i+\overline{\varepsilon}_{i\cdot}-\overline{\varepsilon})^2$$

$$=\sum_{i=1}^{r}m_ia_i^2+\sum_{i=1}^{r}m_i\overline{\varepsilon}_{i\cdot}^2-n\overline{\varepsilon}^2+2\sum_{i=1}^{r}m_ia_i(\overline{\varepsilon}_{i\cdot}-\overline{\varepsilon})$$

又由 $E(\overline{\varepsilon}_{i\cdot})=0$， $E(\overline{\varepsilon})=0$，故

$$E(SSA)=\sum_{i=1}^{r}m_ia_i^2+\sum_{i=1}^{r}m_iE(\overline{\varepsilon}_{i\cdot}^2)-nE(\overline{\varepsilon}^2)=\sum_{i=1}^{r}m_ia_i^2+\sum_{i=1}^{r}m_i\frac{\sigma^2}{m_i}-n\cdot\frac{\sigma^2}{n}$$

$$=\sum_{i=1}^{r}m_ia_i^2+(r-1)\sigma^2$$

从上面的分析过程中可得如下定理。

定理 6-2（平方和的期望定理）　在一个因素的方差分析模型中，有

$$E(SSE)=(n-r)\sigma^2$$

$$E(SSA)=\sum_{i=1}^{r}m_ia_i^2+(r-1)\sigma^2$$

定理 6-3（误差偏差平方和分布定理）　在一个因素的方差分析模型中，有

$$\frac{SSE}{\sigma^2}\sim\chi^2(n-r)$$

定理 6-4（因素 A 的偏差平方和分布定理）　在一个因素的方差分析模型中，当假设 H_0 为真时，有

$$E\left(\frac{SSA}{r-1}\right)=\sigma^2$$

$$\frac{SSA}{\sigma^2}\sim\chi^2(r-1)$$

SSA 与 SSE 相互独立，且 $F=\dfrac{SSA/(r-1)}{SSE/(n-r)}\sim F(r-1,n-r)$。

因此可采用统计量 F 来检验假设 H_0。当 H_0 为假时，分子的均值要比分母的均值大，因而取如下拒绝域

$$W=\{F\geqslant c\}$$

是合理的。对给定的显著性水平 a，c 应满足

$$P\{F\geqslant c\}=a$$

当取 $c=F_{1-a}(r-1,n-r)$ 时，便有 $P\{F\geqslant c\}=a$，故得拒绝域为

$$W=\{F \geqslant F_{1-a}(r-1,n-r)\}$$

通常把以上求统计量的计算列成一张表格，称为方差分析表（见表 6-3），相应的χ^2分布中的自由度也列于表中，偏差平方和与自由度的比称为均方和。



表 6-3　单因素方差分析表

偏差来源	偏差平方和	自由度	均方和	F 值
A	SSA	$f_A=r-1$	$V_A=SSA/f_A$	$F=V_A/V_E$
E	SSE	$f_E=n-r$	$V_E=SSE/f_E$	
T	SST	$f_T=n-1$		

综上所述，单因素方差分析的步骤如下。

1）依次列出第i $(i=1,2,\cdots,r)$个变量对应容量为m_i的样本$y_{i1},y_{i2},\cdots,y_{im_i}$，确定试验中因素的水平数$r$、各水平下的样本容量$m_i$、数据总数$n=\sum_{i=1}^{r}m_i$，同时明确显著性水平$\alpha$。

2）计算各水平下的数据和$y_{i\cdot}=\sum_{j=1}^{m_i}y_{ij}(i=1,2,\cdots,r)$及总和$y_{\cdot\cdot}=\sum_{i=1}^{r}\sum_{j=1}^{m_i}y_{ij}$，计算各数据$y_{ij}$平方之和$\sum_{i=1}^{r}\sum_{j=1}^{m_i}y_{ij}^2$，在此基础上计算$\sum_{i=1}^{r}\frac{y_{i\cdot}^2}{m_i}$，$\frac{y_{\cdot\cdot}^2}{n}$。

3）利用步骤 2）中的结果，计算 SST，SSA 和 SSE。其中

$$SST=\sum_{i=1}^{r}\sum_{j=1}^{m_i}(y_{ij}-\overline{y})^2=\sum_{i=1}^{r}\sum_{j=1}^{m_i}y_{ij}^2-\frac{y_{\cdot\cdot}^2}{n}$$

$$SSA=\sum_{i=1}^{r}m_i(\overline{y}_{i\cdot}-\overline{y})^2=\sum_{i=1}^{r}\frac{y_{i\cdot}^2}{m_i}-\frac{y_{\cdot\cdot}^2}{n}$$

$$SSE=SST-SSA$$

4）确定自由度$f_A=r-1$和$f_E=n-r$，计算各类均方和$V_A=SSA/f_A$和$V_E=SSA/f_E$，求出检验统计量$F=V_A/V_E$，即得到了单因素方差分析表中的各项内容。

5）求出临界值$F_{1-a}(f_A,f_E)$，确定拒绝域$W=\{F \geqslant F_{1-a}(f_A,f_E)\}$。若$F\in W$，则做出拒绝原假设$H_0$的结论；否则，接受$H_0$。

或者由最小显著性概率p做出检验决策，当$p<\alpha$时拒绝原假设。

对于例 6-2，所谓方差分析，即检验假设H_0：$\mu_1=\mu_2=\mu_3$，其中$\mu_i(i=1,2,3)$是第i个变量的均值。

其实现的 MATLAB 程序代码如下：

```
>> %MATLAB 数据处理(1)
clear all;
y=[73 66 89 82 43 80 63 88 78 91 76 85 94 80 96 68 79 71 71 87 68 59 76 80];
r=3;
m1=7;m2=8;m3=9;                    %各总体的样本容量
n=m1+m2+m3;
```

```
alpha=0.05;
y1=sum(y(1:m1));                        %第一种教学方法下学生的成绩之和
y2=sum(y((m1+1):(m1+m2)));              %第二种教学方法下学生的成绩之和
y3=sum(y((m1+m2+1):n));                 %第三种教学方法下学生的成绩之和
y4=sum(y);                              %各学生成绩之和
yy=sum(y.^2);                           %各学生成绩平方之和
g=y1^2/m1+y2^2/m2+y3^2/m3;
SST=yy-y4^2/n;                          %总的偏差平方和
SSA=g-y4^2/n;                           %因素的偏差平方和
SSE=SST-SSA;                            %误差偏差平方和
g1=SSA/(r-1);                           %偏差均方和
g2=SSE/(n-r);                           %误差偏差均方和
FEST=g1/g2;                             %由样本计算出的F值
FLJ=finv(1-alpha,r-1,n-r);              %应用MATLAB统计工具箱中的finv函数求得临界值
p=1-fcdf(FEST,r-1,n-r);
if FEST>FLJ
    h=1;
else
    h=0;
end
alpha,h,p,FEST,FLJ
```

运行程序，输出如下：

```
alpha =      0.0500
h =        1
p =      0.0211
FEST =      4.6638
FLJ =      3.4668
```

计算结果表明，在显著性水平α=0.05 下，h=1、p<alpha（拒绝原假设），即认为 3 种教学方法有显著差异。

6.2.2 效应与误差方差的估计

1. 效应与误差方差的点估计

由模型（6-1）（即式（6-1））知各 y_{ij} 相互独立，因而可用极大似然估计法求出各效应与 σ^2 的估计。不难证明如下定理。

定理 6-5（效应与误差方差的点估计定理）

$$\hat{\mu} = \overline{y},\quad \hat{\mu}_i = \overline{y}_{i\cdot},\quad \overline{a}_i = \overline{y}_{i\cdot} - \overline{y}, i = 1,2,\cdots,r$$

$$\hat{\sigma}_m^2 = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{m_i}(y_{ij} - \overline{y}_{i\cdot})^2 = \frac{SSE}{n}$$

σ^2的无偏差估计是

$$\hat{\sigma}^2 = \frac{SSE}{n-r}$$

证明略。



2. μ_i 的置信水平为 $1-\alpha$ 的置信区间

利用统计量法，可以构造 μ_i 的置信区间。

从 μ_i 的点估计 $\bar{y}_{i\cdot}$ 出发，由于前面已证明 $\bar{y}_{i\cdot} \sim N\left(\mu_i, \frac{\sigma^2}{m_i}\right)$，又有 $\frac{SSE}{\sigma^2} \sim \chi^2(f_E)$，这里 $f_E = n-r$，且 $\bar{y}_{i\cdot}$ 与 SSE 独立，因而可以构造一个服从 t 分布的统计量

$$t_i = \frac{\dfrac{\bar{y}_{i\cdot} - \mu_i}{\dfrac{\sigma}{\sqrt{m_i}}}}{\sqrt{\dfrac{\dfrac{SSE}{\sigma^2}}{f_E}}} = \frac{\bar{y}_{i\cdot} - \mu}{\dfrac{\hat{\sigma}}{\sqrt{m_i}}} \sim t(f_E)$$

因而从

$$P\left\{|t_i| \leqslant t_{1-\frac{\alpha}{2}}(f_E)\right\} = 1-\alpha$$

可得 μ_i 的置信水平为 $1-\alpha$ 的置信区间为

$$\left(\bar{y}_{i\cdot} - t_{1-\frac{\alpha}{2}}(f_E)\frac{\hat{\sigma}}{\sqrt{m_i}}, \bar{y}_{i\cdot} + t_{1-\frac{\alpha}{2}}(f_E)\frac{\hat{\sigma}}{\sqrt{m_i}}\right)$$

这里，$\hat{\sigma} = \sqrt{\dfrac{SSE}{f_E}}$。

【例 6-3】 求例 6-2 中每一种教学方法下学生平均英语成绩的点估计和置信水平为 0.95 的置信区间。

```
>> %MATLAB 数据处理(2)
clear all;
alpha=0.05;
m1=7;m2=8;m3=9;    %各总体的样本容量
n=m1+m2+m3;
r=3;
fE=n-r;
y1=496;              %引用 MATLAB 数据处理(1)中的结果,下同
y2=688;
y3=659;
MU1=y1/m1            %第一种教学方法下学生平均英语成绩的点估计
MU2=y2/m2            %第二种教学方法下学生平均英语成绩的点估计
MU3=y3/m3            %第三种教学方法下学生平均英语成绩的点估计
T=tinv(1-alpha/2,fE);
SSE=2.3404e+003;            %引用 MATLAB 数据处理(1)中的结果
SIGMA=sqrt(SSE/(n-r));    %英语成绩标准差的无偏估计
a=[MU1-T*SIGMA/sqrt(m1),MU1+T*SIGMA/sqrt(m1)];
b=[MU2-T*SIGMA/sqrt(m2),MU2+T*SIGMA/sqrt(m2)];
c=[MU3-T*SIGMA/sqrt(m3),MU3+T*SIGMA/sqrt(m3)];
a,b,c     %3 种教学方法下平均英语成绩的置信区间
```

运行程序，输出如下：

```
MU1 =    70.8571
MU2 =     86
MU3 =    73.2222
a =
    62.5592    79.1551
b =
    78.2380    93.7620
c =
    65.9041    80.5403
```

计算结果表明，3 种教学方法下学生的英语成绩分别为 70.8571、86、73.2222；置信水平为 0.95 的置信区间分别为[62.5592 79.1551]，[78.2380 93.7620]，[65.9041 80.5403]。

6.2.3 重复数相同的方差分析

当在因素 A 的每一水平下重复试验次数相同，即当 $m_1 = m_2 = \cdots = m_r$ 时，上述一些表达式可以简化。若记每一水平下重复次数为 m，则效应约束条件可简化为

$$\sum_{i=1}^{r} a_i = 0$$

SSA 的计算公式可简化为

$$SSA = \frac{1}{m}\sum_{i=1}^{r} y_{i\cdot}^2 - \frac{y_{\cdot\cdot}^2}{n}$$

μ_i 的置信水平为 $1-\alpha$ 的置信区间可改为

$$\left(\overline{y}_{i\cdot} - t_{1-\frac{\alpha}{2}}(f_E)\frac{\hat{\sigma}}{\sqrt{m}}, \overline{y}_{i\cdot} + t_{1-\frac{\alpha}{2}}(f_E)\frac{\hat{\sigma}}{\sqrt{m}}\right)$$

其他一切都不变。对于重复数相同的单因素方差分析，MATLAB 提供了 anova1 函数来处理单因素方差分析的问题。anova1 函数主要是比较多组数据的均值，然后返回这些均值相等的概率，从而判断这一因素是否对试验指标有显著影响。

其调用格式如下：

```
p = anova1(X)
p = anova1(X,group)
p = anova1(X,group,'displayopt')
[p,table] = anova1(...)
[p,table,stats] = anova1(...)
```

其中，p = anova1(X)对样本 X 中的两列或多列数据进行均衡的单因素方差分析，以比较各列的均值。函数返回“零假设”（即 X 中各列的均值相同）成立的概率值。如果概率值接近于零，则零假设值得怀疑，表明各列的均值事实上是不同的。p = anova1(X,group)对样本 X 中由矢量 group 索引的两组或多组数据进行单因素方差分析以比较各列的均值。输入参数 group 标明矢量 X 中相应元素的组别。group 中的值为整数，最大值为需要比较的不同组的数量，最小值为 1。每组至少应有一个元素，但并不要求每组的元素个数相同，因此适合于



数据不均衡的情况。用于决定结果是否具有统计上的显著性的概率值大小限制的选择留给用户。[p,table,stats] = anova1(...)anova1 同时还显示一张表 table 和一幅图 stats。表为标准的 ANOVA 表，表中将 *X* 中数据的变化分别分成两部分：

① 由各列均值的差异而产生的变化。

② 由各列的数据及其均值间的差异而导致的变化。

ANOVA 表至少具有 5 列数据。

① 第一列标明数据源。

② 第二列给出数据源的均方和（*SS*）。

③ 第三列给出相应数据源的自由度 *df*。

④ 第四列给出均方值 *p*，即比率 *SS*/*df*。

⑤ 第五列给出 *F* 统计量。

p 值是 *F* 的函数（fcdf）。随着 *F* 的增加，*p* 值减小。在 box 图中，各列数据的图的中心线若表现出较大差异，则相应于 *F* 值较大以及 *p* 值较小。

【例 6-4】 *X* 中的 5 列数据分别为 1~5 的常数与均值为 0、标准差为 1 的正态随机干扰量之和。

其实现的 MATLAB 程序代码如下（结果见图 6-1）：

```
>> X=meshgrid(1:5)
X =
      1      2      3      4      5
      1      2      3      4      5
      1      2      3      4      5
      1      2      3      4      5
      1      2      3      4      5
>> X=X+normrnd(0,1,5,5)
X =
    0.5674    3.1909    2.8133    4.1139    5.2944
   -0.6656    3.1892    3.7258    5.0668    3.6638
    1.1253    1.9624    2.4117    4.0593    5.7143
    1.2877    2.3273    5.1832    3.9044    6.6236
   -0.1465    2.1746    2.8636    3.1677    4.3082
>> p=anova1(X)
p =   1.2765e-006
```

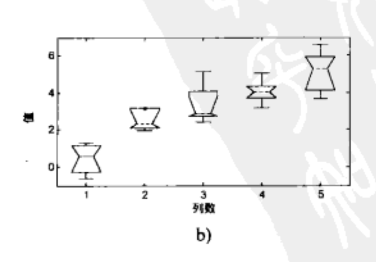
Figure 1: One-way ANOVA

File Edit View Insert Tools Desktop Window Help

ANOVA Table

Source	SS	df	MS	F	Prob>F
Columns	62.4487	4	15.6122	19.18	1.27648e-006
Error	16.2792	20	0.814		
Total	78.7279	24			

a)



b)

图 6-1 单因素方差分析

a) ANOVA 效果表 b) box 效果图

由计算结果，观察所提供的 X 的随机数据样本，可知其各列均值相同的概率小于 $6/10^5$。

【例 6-5】 某钢厂检查一月上旬的 5 天中生产的钢锭质量，结果见表 6-4。

表 6-4 某钢厂生产的钢锭质量 （单位：kg）

日 期	质 量			
1	5500	5800	5740	5710
2	5440	5680	5240	5600
4	5440	5410	5430	5400
9	5640	5700	5660	5700
10	5610	5700	5610	5400

试检验不同日期生产的钢锭有无显著差异（$\alpha = 0.05$）。

分析：把不同日期生产的钢锭质量分别看做一个变量。检验它们的平均质量是否有明显差异相当于比较 5 个变量的均值是否一致。假定：①5 个变量均服从正态分布。②每一变量的方差相同。③从 5 个变量抽取的样本相互独立。采用方差分析法来检验不同日期生产的钢锭质量是否有明显差异。

设第 i 个变量的均值为 μ_i，假设不同日期生产的钢锭平均质量无显著差异，那么就要检验如下假设：

$$H_0: \quad \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$$

其实现的 MATLAB 程序代码如下：

```
>> clear;
A1=[5500 5800 5740 5710]';
A2=[5440 5680 5240 5600]';
A3=[5400 5410 5430 5400]';
A4=[5640 5700 5660 5700]';
A5=[5610 5700 5610 5400]';
X=[A1,A2,A3,A4,A5];
[p,table,stats]=anova1(X,[],'on')
```

运行程序，输出如下（见图 6-2）：

```
p =      0.0220
table =
    'Source'    'SS'        'df'    'MS'              'F'         'Prob>F'
    'Columns'   [227680]    [ 4]    [       56920]    [3.9496]    [0.0220]
    'Error'     [216175]    [15]    [1.4412e+004]          []           []
    'Total'     [443855]    [19]               []          []           []
stats =
    gnames: [5x1 char]
         n: [4 4 4 4 4]
    source: 'anova1'
     means: [5.6875e+003 5490 5410 5675 5580]
```

df: 15
 s: 120.0486

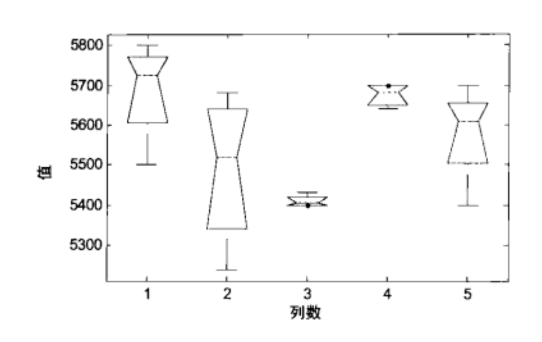




图 6-2　5 天生产钢锭质量的 box 图

结果表明：①返回值 p=0.0220<0.05，认为不同日期生产的钢锭平均质量有显著差异。②方差分析表（table）中有 6 列，第一列声明 X 中可变化性的来源；第二列显示平方和；第三列显示与每一种可变化有关的自由度；第四列显示第二列数据与第三列数据的比值；第五列显示 F 统计量数据值，是第四列数据的比值；第六列显示检验的最小显著性概率，即第一列输出参数值。③stats 返回的附加统计数据结构中 means 一行给出了各日生产的钢锭平均质量的点估计。④从方差分析 box 图容易看出不同日期生产的钢锭平均质量之间的直观差异。

6.2.4 多重比较

若检验结果拒绝了 H_0，进一步分析哪些水平之间的差异是显著的、哪些水平对实验结果的影响最大、哪些水平次之，这在实际应用中往往是很重要的。此项工作通常称为均值的多重比较。

对任意两个水平均值之间有无显著差异进行多重比较，即同时检验以下$\binom{r}{2}$个假设：

$$H_0^{ij}:\ \mu_i=\mu_j,\quad H_1^{ij}:\ \mu_i\neq\mu_j,\quad i<j;i,j=1,2,\cdots,r$$

检验的统计量为

$$t=\frac{(\overline{y}_{i\cdot}-\overline{y}_{j\cdot})}{\sqrt{s^2\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}}$$

其中，$s^2=\frac{SSE}{n-r}$。对于 H_0^{ij} 的检验水平 α'，当 $|t|>t_{1-\frac{\alpha'}{2}}(n-r)$ 时拒绝 H_0^{ij}。或等价地，当置信水平为 $100(1-\alpha')\%$ 的 $\mu_i-\mu_j$ 置信区间

$$t=(\overline{y}_{i\cdot}-\overline{y}_{j\cdot})\pm t_{1-\frac{\alpha'}{2}}(n-r)\cdot s\cdot\sqrt{\frac{1}{n_i}+\frac{1}{n_j}}$$

不包含 0 时拒绝 H_0^{ij}，从而拒绝 H_0。

由于多重比较所进行的一系列检验均构成对于假设的检验，因此要使得所有检验犯第一

类错误的总概率不超过给定的α，就需要选取适当的α'。检验H_0和检验H_0^{ij}的交$\bigcap_{1\leqslant i\leqslant j\leqslant r} H_0^{ij}$等价于：当所有的$H_0^{ij}$成立时，$H_0$必成立，反之亦然。以$A_{ij}$记为$H_0^{ij}$的拒绝域，则

$$P\{拒绝H_0 \mid H_0\} = P\{至少有一个A_{ij}发生 \mid H_0\}$$
$$= P\{A_{12} + A_{13} + \cdots + A_{r,r-1} \mid H_0\} \leqslant \sum_{1\leqslant i\leqslant j\leqslant r} P\{A_{ij} \mid H_0\}$$
$$\leqslant \sum_{1\leqslant i\leqslant j\leqslant r} P\{A_{ij} \mid H_0^{ij}\} \leqslant \binom{r}{2}\alpha'$$

要使犯第一类错误的总概率$P\{拒绝H_0 \mid H_0\} \leqslant \alpha'$，只要取$\alpha' = \alpha \Big/ \binom{r}{2}$。

通过$\binom{r}{2}$个均值比较，检验假设H_0的优点是它不仅可知$\mu_1, \mu_2, \cdots, \mu_r$有差别，而且知道差别在哪。但此方法计算量大，同时由于要保证总的检验水平，α'取得比较小，从而一般来说，比起直接应用方差分析，增大了犯第二类错误的概率，这意味着可能会出现这样的情形：用F检验结果是显著的，但用两两比较没有任何两个水平有显著差异。下面的 LSD 方法在某种程度上可以弥补这个缺陷，但真实水平是近似的。

LSD 方法是由 R.A.Fisher 提出，又经过后人修正的。方法如下：

① 给定检验水平α，用方差分析法检验H_0。

② 如果拒绝H_0，则继续比较水平之间的差异，否则停止。

③ 对于水平i, j，μ_i与μ_j的最小显著差异为

$$LSD_{ij} = t_{1-\frac{\alpha}{2}}(n-r)\sqrt{s^2\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

④ 当$\left|\overline{y}_{i\cdot} - \overline{y}_{j\cdot}\right| \geqslant LSD_{ij}$时，认为$\mu_i$与$\mu_j$不同。

【例 6-6】 用多重比较的方法确定例 6-2 中哪些水平之间的差异是显著的，同时确定使学生的平均英语成绩最高的教学方法。

分析：在例 6-2 中，已经得出 3 种教学方法有显著性差异，即教学方法这一因素对学生的英语成绩是有显著影响的。进一步分析到底哪两种教学方法对学生的成绩影响差异显著，就需要对 3 个变量进行多重比较了。多重比较的方法很多，按照上面介绍的 LSD 方法，利用 MATLAB 计算如下。

```
%MATLAB 数据处理(3)
>> %MATLAB 数据处理(2)
clear all;
alpha=0.05;
m1=7;m2=8;m3=9;   %各总体的样本容量
n=m1+m2+m3;
r=3;
t=tinv(1-alpha/2,n-r);
```



```
SSE=2.3404e+003;  %引用 MATLAB 数据处理(1)中的结果
LSD12=t*sqrt(SSE/(n-r))*sqrt(1/m1+1/m2);
LSD13=t*sqrt(SSE/(n-r))*sqrt(1/m1+1/m3);
LSD23=t*sqrt(SSE/(n-r))*sqrt(1/m2+1/m3);
MU1=70.8571;  %引用 MATLAB 数据处理(2)中的结果,下同
MU2=86;
MU3=55.1111;
if abs(MU1-MU2)>LSD12
    h(1)=1;
else
    h(1)=0;
end
if abs(MU1-MU3)>=LSD13
    h(2)=1;
else
    h(2)=0;
end
if abs(MU2-MU3)>=LSD23
    h(3)=1;
else
    h(3)=0;
end
h  %结果,依次显示第一和第二,第一和第三,第二和第三种方法下，学生平均成绩差异的显著性
```

运行程序，输出如下：

```
h =
     1     1     1
```

计算结果表明：3 种教学方法对学生的英语平均成绩的影响有显著差异；第二种教学方法使学生的英语平均成绩最高。

6.2.5 方差齐性检验

在单因素方差分析中，假定 r 个不同水平下的响应变量 y_i 服从 $N(\mu_i,\sigma_i^2)$ $(i=1,2,\cdots,r)$，并要求这 r 个正态变量的方差相等，这一要求简称为方差齐性。一般而言，实际应用中在进行方差分析之前，有两项预备性分析是不可缺少的。一是这 r 个变量的正态性检验（检验方法在第 5 章已经介绍）；二是这 r 个正态变量的方差齐性检验。

方差齐性检验的假设为

$$H_0:\ \sigma_1^2=\sigma_2^2=\cdots=\sigma_r^2;\quad H_1:\ \sigma_1^2=\sigma_2^2=\cdots=\sigma_r^2\text{ 不全相等}$$

备择假设往往略去不写。

方差齐性通常采用 Bartlett 检验方法。下面简单介绍 Bartlett 检验的基本思路和检验统计量的构造。

设第 i 个变量抽取了容量为 m_i 的样本 $y_{i1},y_{i2},\cdots,y_{im_i}$，其样本方差为

$$s_i^2 = \frac{1}{m_i}\sum_{j=1}^{m_i}(y_{ij} - \overline{y}_{i\cdot})^2 = \frac{Q_i}{f_i}, \quad i = 1,2,\cdots,r$$

其中，$Q_i = \sum_{i=1}^{m_i}(y_{ij} - \overline{y}_i)$，$f_i = m_i - 1$ 分别为该变量的样本偏差平方和与自由度。于是，随机误差均方和

$$MSSE = \frac{1}{f_E}SSE = \frac{1}{f_E}\sum_{i=1}^{r}Q_i = \sum_{i=1}^{r}\frac{f_i}{f_E}s_i^2$$

是 r 个变量样本方差 $s_i^2(i=1,2,\cdots,r)$ 的加权算术平均数。又令

$$GMSSE = \left[\prod_{i=1}^{r}\left(s_i^2\right)^{f_i}\right]^{\frac{1}{f_E}}$$

是 r 个变量样本方差 $s_i^2\left(i=1,2,\cdots,r\right)$ 的几何平均数，$f_E = \sum_{i=1}^{r}f_i$。

由于恒有 $GMSSE \leqslant MSSE$，并且等号成立的充分必要条件是 $s_1^2 = s_2^2 = \cdots = s_r^2$，所以，诸样本方差 $s_i^2\left(i=1,2,\cdots,r\right)$ 间的差异越大，$GMSSE$ 和 $MSSE$ 的差异越大。换句话说，当 H_0 为真时，比值 $GMSSE/MSSE$ 接近于 1。反之，比值 $GMSSE/MSSE$ 比较大时，H_0 值得怀疑。这个结论对 $\ln(GMSSE/MSSE)$ 也成立。于是，H_0 的拒绝域应有如下形式：

$$W = \{\ln(\ln(GMSSE/MSSE) \geqslant d\}$$

Bartlett 证明了，在大样本条件下

$$B = \frac{f_E}{c}(\ln MSSE - \ln GMSSE) \sim \chi^2(r-1)$$

其中，$c = 1 + \frac{1}{3(r-1)}\left(\sum_{i=1}^{r}\frac{1}{f_i} - \frac{f}{f_E}\right)$。显然，一般情况下 $c > 1$。

通常，当各个变量的样本容量 $m_i \geqslant 5(i=1,2,\cdots,r)$ 时，也可以用统计量 B 作为 H_0 的检验统计量。在显著性水平 α 下，拒绝域为

$$W = \left\{B \geqslant \chi_{1-\alpha}^2(r-1)\right\}$$

实际计算时，检验统计量采用

$$B = \frac{1}{c}\left(f_E\ln\left(SSE/f_E\right) - \sum_{i=1}^{r}f_i\ln s_i^2\right)$$

的形式更方便一些。

【例 6-7】 对例 6-2 中 3 种教学方法下学生的英语成绩这 3 个变量作方差齐性检验。

分析：假设 H_0：$\sigma_1^2 = \sigma_2^2 = \sigma_3^2$，即 3 个变量的方差相等。按照上述结论，分别求得例 6-2 中检验统计量 B 的值和本题的拒绝域，经过比较得出结论。

```
>> %MATLAB 数据处理(4)
clear all;
```



```
y=[73 66 89 82 43 80 63 88 78 91 76 85 94 80 96 68 79 71 71 87 68 59 76 80];
alpha=0.05;
m1=7;m2=8;m3=9;  %各总体的样本容量
n=m1+m2+m3;
r=3;
SSE=2.3404e+003;  %引用 MATLAB 数据处理(1)中的结果
n=m1+m2+m3;
fE=n-r;
c=(1/(m1-1)+1/(m2-1)+1/(m3-1)-1/fE)/(3*(r-1))+1;
s1=var(y(1:m1));
s2=var(y((m1+1):(m1+m2)));
s3=var(y((n-m3+1):n));
chi2EST=(fE*log(SSE/fE)-(m1-1)*log(s1)-(m2-1)*log(s2)-(m3-1)*log(s3))/c;
LJZ=chi2cdf(1-alpha,r-1);
p=1-chi2cdf(chi2EST,r-1);
if chi2EST>LJZ
    h=1;
else
    h=0;
end
alpha,h,p,chi2EST,LJZ
```

运行程序，输出如下：

```
alpha =     0.0500
h =     1
p =     0.1330
chi2EST =     4.0348
LJZ =     0.3781
```

计算结果表明，在显著性水平α=0.05 下，h=1、p>alpha 不能拒绝原假设，即认为 3 种教学方法下学生的英语成绩这 3 个变量的方差相等。

下面，对单因素方差分析的应用步骤总结如下。

① 对各个变量（不同的因素水平）的正态性进行检验。

② 对各个变量的方差齐性进行检验（如例 6-2 中的 MATLAB 数据处理（4））。

③ 当各个变量的正态性和方差齐性得到检验后，进行方差分析（如例 6-2 中的 MATLAB 数据处理（1））。在各个变量的正态性和方差齐性没有得到验证的情况下，严格地说不宜再作方差分析。但是，有关研究表明方差分析的 F 统计量有较好的稳定性，即使正态性和方差齐性没有得到验证，也可以进行粗略的方差分析以供参考。

④ 在方差分析拒绝各个变量的均值一致的原假设后，应进行多重比较（如例 6-2 中的 MATLAB 数据处理（3））。

⑤ 无论方差分析是否拒绝原假设，都应对每个变量的均值作出估计（如例 6-2 中的 MATLAB 数据处理（2））。

6.3 双因素方差分析

上面讲解了单因素实验的方差分析问题，但在科研和生产实践中，常常需要同时研究两个以上因素对实验结果的影响情况。若同时研究两个因素对实验结果的影响，例如，研究不同浸提温度和浸提时间对茶叶有效成分提取的影响，就要对两个实验因素进行方差分析。对于双因素方差分析，其基本思想和方法与单因素方差分析相似，前提条件仍然是要满足独立，方差具有齐性、正态。不同的是，在双因素实验中，有可能出现交互作用。按照是否进行重复实验，双因素方差分析又分为两种，下面分别给予介绍。

6.3.1 双因素无重复实验的方差分析

1. 问题的一般提法

某项实验要同时考察因素 A 和 B 对实验结果的影响，因此 A 取 $A_1,A_2,\cdots,A_a$ 共 a 个水平，因素 B 取 $B_1,B_2,\cdots,B_b$ 共 b 个水平。A 和 B 两因素的每种水平搭配 A_iB_j $(i=1,2,\cdots,a;$ $j=1,2,\cdots,b)$ 各进行一次独立实验，共进行 $a\times b=n$ 次实验，实验数据为 $x_{ij}\,(i=1,2,\cdots,a;j=1,2,\cdots,b)$，这 n 个实验数据见表 6-5。

表 6-5 双因素无重复实验的数据及计算表

因素 A	因素 B						$x_{i\cdot}$	$\bar{x}_{i\cdot}$
	B_1	B_2	$\cdots$	B_j	$\cdots$	B_b		
A_1	x_{11}	x_{12}	$\cdots$	x_{1j}	$\cdots$	x_{1b}	$x_{1\cdot}$	$\bar{x}_{1\cdot}$
A_2	x_{21}	x_{22}	$\cdots$	x_{2j}	$\cdots$	x_{2b}	$x_{2\cdot}$	$\bar{x}_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$	$\vdots$
A_i	x_{i1}	x_{i2}	$\cdots$	x_{ij}	$\cdots$	x_{ib}	$x_{i\cdot}$	$\bar{x}_{i\cdot}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$	$\vdots$
A_a	x_{a1}	x_{a2}	$\cdots$	x_{aj}	$\cdots$	x_{ab}	$x_{a\cdot}$	$\bar{x}_{a\cdot}$
$x_{\cdot j}$	$x_{\cdot 1}$	$x_{\cdot 2}$	$\cdots$	$x_{\cdot j}$	$\cdots$	$x_{\cdot b}$	$x_{\cdot\cdot}$	$\bar{x}_{\cdot\cdot}$

$$x_{i\cdot}=\sum_{j=1}^{b}x_{ij}\,(i=1,2,\cdots,a),\quad \bar{x}_{i\cdot}=\frac{1}{b}x_{i\cdot}$$

$$x_{\cdot j}=\sum_{i=1}^{a}x_{ij}\,(j=1,2,\cdots,b),\quad \bar{x}_{\cdot j}=\frac{1}{a}x_{\cdot j}$$

$$x_{\cdot\cdot}=\sum_{i=1}^{a}\sum_{j=1}^{b}x_{ij},\quad \bar{x}_{\cdot\cdot}=\frac{1}{ab}x_{\cdot\cdot}=\frac{1}{n}x_{\cdot\cdot\cdot}$$

要求分别检验因素 A，B 对实验结果有无显著影响，即检验假设

H_{01}：因素A无显著影响

H_{02}：因素B无显著影响

2. 双因素无重复实验的方差分析步骤

（1）偏差平方和的分解



为了构造检验统计量，仿照单因素方差分析方法，先对偏差平方和进行分解。

$$\begin{aligned}S_T &= \sum_{i=1}^{a}\sum_{j=1}^{b}\left(x_{ij}-\bar{x}_{\cdot\cdot}\right)^2 \\ &= \sum_{i=1}^{a}\sum_{j=1}^{b}\left(\bar{x}_{i\cdot}-\bar{x}_{\cdot\cdot}\right)^2+\sum_{i=1}^{a}\sum_{j=1}^{b}\left(\bar{x}_{\cdot j}-\bar{x}_{\cdot\cdot}\right)^2+\sum_{i=1}^{a}\sum_{j=1}^{b}\left(x_{ij}-\bar{x}_{i\cdot}-\bar{x}_{\cdot j}+\bar{x}_{\cdot\cdot}\right)^2 \\ &= b\sum_{i=1}^{a}\left(\bar{x}_{i\cdot}-\bar{x}_{\cdot\cdot}\right)^2+a\sum_{j=1}^{b}\left(\bar{x}_{\cdot j}-\bar{x}_{\cdot\cdot}\right)^2+\sum_{i=1}^{a}\sum_{j=1}^{b}\left(x_{ij}-\bar{x}_{i\cdot}-\bar{x}_{\cdot j}+\bar{x}_{\cdot\cdot}\right)^2\end{aligned} \tag{6-3}$$

令

$$S_A = b\sum_{i=1}^{a}\left(\bar{x}_{i\cdot}-\bar{x}_{\cdot\cdot}\right)^2 \tag{6-4}$$

S_A 为因素 A 各水平间，即各行间的偏差平方和，反映了因素 A 对实验结果的影响。

令

$$S_B = a\sum_{j=1}^{b}\left(\bar{x}_{\cdot j}-\bar{x}_{\cdot\cdot}\right)^2 \tag{6-5}$$

S_B 为因素 B 各水平间，即各列间的偏差平方和，反映了因素 B 对实验结果的影响。

令

$$S_e = \sum_{i=1}^{a}\sum_{j=1}^{b}\left(x_{ij}-\bar{x}_{i\cdot}-\bar{x}_{\cdot j}+\bar{x}_{\cdot\cdot}\right)^2 \tag{6-6}$$

S_e 为误差偏差平方和，即组内偏差平方和，反映了实验误差的大小。

于是式（6-3）可记为

$$S_T = S_A + S_B + S_e \tag{6-7}$$

（2）偏差平方和的简化计算

$$S_T = \sum_{i=1}^{a}\sum_{j=1}^{b}\left(x_{ij}-\bar{x}_{\cdot\cdot}\right)^2 = \sum_{i=1}^{a}\sum_{j=1}^{b}x_{ij}^2-\frac{1}{n}x_{\cdot\cdot}^2 = Q_T - C_T \tag{6-8}$$

$$S_A = b\sum_{i=1}^{a}\left(\bar{x}_{i\cdot}-\bar{x}_{\cdot\cdot}\right)^2 = \frac{1}{b}\sum_{i=1}^{a}x_{i\cdot}^2-\frac{1}{n}x_{\cdot\cdot}^2 = Q_A - C_T \tag{6-9}$$

$$S_B = a\sum_{j=1}^{b}\left(\bar{x}_{\cdot j}-\bar{x}_{\cdot\cdot}\right)^2 = \frac{1}{a}\sum_{j=1}^{b}x_{\cdot j}^2-\frac{1}{n}x_{\cdot\cdot}^2 = Q_B - C_T \tag{6-10}$$

$$S_e = S_T - S_A - S_B \tag{6-11}$$

（3）计算自由度和方差

S_T 的自由度：$f_T = ab-1 = n-1$

S_A的自由度：$f_A = a-1$

S_B的自由度：$f_B = b-1$

S_e的自由度：$f_e = f_T - f_A - f_B = (a-1)(b-1)$

将各偏差平方和除以相应的自由度，可求得各行间、各列间和误差的方差如下：

行间方差

$$V_A = \frac{S_A}{f_A} = \frac{S_A}{a-1} \tag{6-12}$$

列间方差

$$V_B = \frac{S_B}{f_B} = \frac{S_B}{b-1} \tag{6-13}$$

误差方差

$$V_e = \frac{S_e}{f_e} = \frac{S_e}{(a-1)(b-1)} \tag{6-14}$$

（4）显著性检验

数学上可以证明：假设H_{01}为真时，统计量

$$F_A = \frac{V_A}{V_e} = \frac{S_A/(a-1)}{S_e/(a-1)(b-1)} \sim F\left[(a-1),(a-1)(b-1)\right] \tag{6-15}$$

假设H_{02}为真时，统计量

$$F_B = \frac{V_B}{V_e} = \frac{S_B/(b-1)}{S_e/(a-1)(b-1)} \sim F\left[(b-1),(a-1)(b-1)\right] \tag{6-16}$$

因此，利用F_A与F_B就可以分别对因素A和B作用的显著性进行检验。对于给定的显著性水平α，在相应的自由度下查出$F_{A,\alpha}$和$F_{B,\alpha}$，若$F_A \geqslant F_{A,\alpha}$，拒绝$H_{01}$，反之，则接受$H_{01}$；若$F_B \geqslant F_{B,\alpha}$，则拒绝$H_{02}$，反之，则接受$H_{02}$。

6.3.2 双因素重复实验的方差分析

求解双因素方差分析问题的 MATLAB 统计学工具箱函数为 anova2。

其调用格式如下：

```
P = anova2(X, reps)
p = anova2(X,reps,'displayopt')
[p,table] = anova2(...)
[p,table,stats] = anova2(...)
```

其中，双因素方差分析是一种两因素、多水平析因试验数据的统计分析方法。其目的在于确认来自不同组的数据是否具有相同的均值。

假设一个汽车制造公司有两个工厂，都分别制造 3 种汽车。下面来考察汽车的燃气里程（即每升汽油所跑里程数）随汽车种类和工厂不同而变化的情况。由于工厂制造方

法的差异，使燃气里程有总体的差别；由于设计规定的差异，不同种类汽车的燃气里程也可能不同。另外，制造方法和设计规定二者之间也可能存在综合效应，从而影响汽车的燃气里程。因此，除非对工厂和汽车种类相结合进行观察，否则不可能观测到交互作用。



双因素方差分析是处理这种问题的典型方法。首先建立问题的数学模型：

$$y_{ijk} = \mu + \alpha_j + \beta_i + \gamma_{ij} + \varepsilon_{ijk}$$

式中，y_{ijk} 是观测值矩阵；μ 是样本总均值（常数均值）；α_j 是列元素为组均值的矩阵（各行 α 的总和为 0）；β_i 是行元素为组均值的矩阵（各列 β 的总和为 0）；γ_{ij} 是交互作用项（矩阵）（各行、各列 γ 的总和为 0）；ε_{ijk} 是随机干扰矩阵。

返回“零假设”（即列数据的均值与行数据的均值相同）成立的概率值 p。如果概率值接近于零，则假设值得怀疑。用于决定结果是否有统计上的显著性的概率值限制的选择留给用户。通常认为，如果 p 值小于 0.05 或 0.1，则结果较显著，同时也显示一个标准方差分析表（ANOVA 表）。其中，按照 reps 参数值将 x 中数据的变化情况分成 3 部分或 4 部分。

① 由各列均值差异而产生的变化。

② 由各行均值差异而产生的变化。

③ 由列和行因素的交互作用而导致的变化（如果 reps 值大于其默认值 1）。

④ 其他因素。

ANOVA 表共有 5 列数据：

① 第一列标明数据源。

② 第二列给出相应数据源的均方和（SS）。

③ 第三列给出相应数据源的自由度 df。

④ 第四列给出均方值，即比率 ss/df。

⑤ 第五列给出 F 统计量，即均方比。

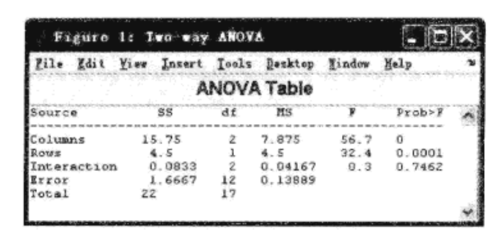
p 值是 F 的函数（fcdf）。随着 F 值的增加，p 值减少。

【例 6-8】 双因素方差分析。

```
>> load popcorn
popcorn
p=anova2(popcorn,3)
```

运行程序，输出如下（效果见图 6-3）：

```
popcorn =
    5.5000    4.5000    3.5000
    5.5000    4.5000    4.0000
    6.0000    4.0000    3.0000
    6.5000    5.0000    4.0000
    7.0000    5.5000    5.0000
    7.0000    5.0000    4.5000
p =
    0.0000    0.0001    0.7462
```



Source	SS	df	MS	F	Prob>F
Columns	15.75	2	7.875	56.7	0
Rows	4.5	1	4.5	32.4	0.0001
Interaction	0.0833	2	0.04167	0.3	0.7462
Error	1.6667	12	0.13889		
Total	22	17			

图 6-3　双因素方差分析

【例 6-9】 为了考察固化时间及固化温度对胶黏剂粘接材料强度的影响，进行了 12 次试验之后得到的结果见表 6-6，要求分析固化时间和固化温度的不同是否对粘接强度有显著影响。

表 6-6　不同固化时间、温度下的粘接强度

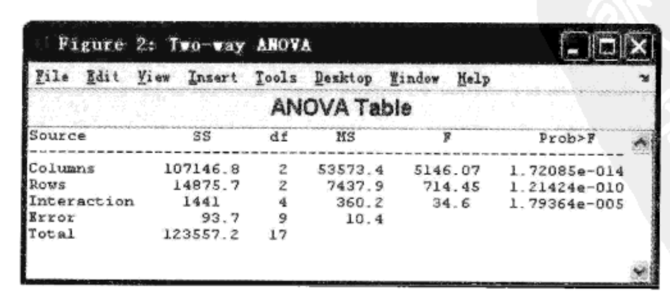
时间/s	温　度		
	25°C	50°C	90°C
10	52.3	136.8	230.5
	58.9	132.1	224.8
30	83.6	157.3	260.4
	85.3	153.4	264.8
60	115.6	187.9	323.8
	112.9	185.2	329.9

其实现的 MATLAB 程序代码如下：

```
>> x=[52.3 58.9 83.6 85.3 115.6 112.9;
   136.8 132.1 157.3 153.4 187.9 185.2;
   230.5 224.8 260.4 264.8 323.8 329.9]';
anova2(x,2)
```

运行程序，输出如下（效果见图 6-4）：

```
ans =
  1.0e-004 *
    0.0000      0.0000      0.1794
```



Source	SS	df	MS	F	Prob>F
Columns	107146.8	2	53573.4	5146.07	1.72085e-014
Rows	14875.7	2	7437.9	714.45	1.21424e-010
Interaction	1441	4	360.2	34.6	1.79364e-005
Error	93.7	9	10.4		
Total	123557.2	17			

图 6-4　方差分析表



6.3.3 多因素方差分析

多因素方差分析可以用于确定根据多个因素划分的不同组数据的均值是否不同。如果它们不同，还可以进一步确定这种差异是由哪一个或几个因素引起的。

多因素方差分析是两因素方差分析的一般形式。对 3 个因素的情况，其模型表达式为：

$$y_{ijkl}=\mu+\alpha_{\cdot j\cdot}+\beta_{i\cdot\cdot}+\gamma_{\cdot\cdot k}+(\alpha\beta)_{ij\cdot}+(\alpha\gamma)_{i\cdot k}+(\beta\gamma)_{\cdot jk}+(\alpha\beta\gamma)_{ijk}+\varepsilon_{ijkl}$$

式中，两个连在一起的标记（如$(\alpha\beta)_{ij\cdot}$）表示两个因素之间的交互作用，参数$(\alpha\beta\gamma)_{ijk}$表示 3 个因素之间的交互作用。

MATLAB 统计工具箱实现多因素方差分析的函数为 anovan。

其调用格式如下：

```
p = anovan(X,group)
p = anovan(X,group, 'model')
p = anovan(X,group,'model',sstype)
p = anovan(X,group,'model',sstype, gnames)
p = anovan(X,group,'mode',sstype,gnames,'displayopt')
[p,table]=anovan(...)
[p,table,stats] = anovan(...)
[p,table,stats,terms] = anovan(...)
```

anovan 函数用于实现多因素方差分析。其中，***X*** 是分析的数据矩阵；group 是组的索引；'model'是模型的类型，'model= linear '表示仅仅计算 *N* 个因素的假设检验，'model=interaction'表示计算 *N* 个因素及任意两个因素之间的假设检验，'model=full'表示计算 *N* 个因素及不同水平之间的假设检验；sstype 是平方和的类型；'displayopt=on'显示 ANOVA 表和图，'displayopt=off '则不显示；*p* 返回假设检验结果；table 返回 ANOVA 表；stats 返回一个结构，可用于进一步的多比较分析；terms 返回输出向量的编码。

其相关函数有：anvoal、anova2、multcompare。

例如，在 MATLAB 命令行输入：

```
>>y = [52.7 57.5 45.9 44.5 53.0 57.0 45.9 44.0]';
g1 = [1 2 1 2 1 2 1 2];
g2 = {'hi';'hi';'lo';'lo';'hi';'hi';'lo';'lo'};
g3 = {'may'; 'may'; 'may'; 'may'; 'june'; 'june'; 'june'; 'june'};
p = anovan(y, {g1 g2 g3})
```

运行程序，输出如下（效果见图 6-5）：

```
p =
    0.4174
    0.0028
    0.9140
```

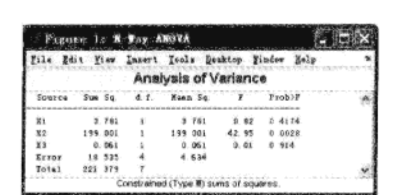


图 6-5　anvoan 函数效果示例

【例 6-10】 分析 3 个因素：出产地（A：欧洲、日本或美国）、是否为四缸的（B）及时间（C）对汽车里程的影响是否显著。

其实现的 MATLAB 程序代码如下：

```
>> clear all;
%装载数据
load carbig
whos
%3 个因素
factornames={'Origin','4Cyl','MfgDate'};
%多因素方差分析
[p,tbl,stats,termvec]=anovan(MPG,{org cyl4 when},2,3,factornames);
p,termvec
```

运行程序，输出如下：

```
Name            Size          Bytes   Class    Attributes
Acceleration    406x1          3248   double
Cylinders       406x1          3248   double
Displacement    406x1          3248   double
Horsepower      406x1          3248   double
MPG             406x1          3248   double
Model           406x36        29232   char
Model_Year      406x1          3248   double
Origin          406x7          5684   char
Weight          406x1          3248   double
cyl4            406x5          4060   char
org             406x7          5684   char
when            406x5          4060   char
```

假设结果为：

```
p =
    0.0000
         0
         0
    0.6422
    0.0001
    0.3348
```

输出向量的编码为：



```
termvec =
     1     0     0
     0     1     0
     0     0     1
     1     1     0
     1     0     1
     0     1     1
```

第一行表示 p 的第一个值对应第一个因素影响的假设检验，第二行表示 p 的第二个值对应第二个因素影响的假设检验，第三行表示 p 的第三个值对应第三个因素影响的假设检验，第四行表示 p 的第四个值对应第一个和第二个因素相互作用影响的假设检验，第五行表示 p 的第五个值对应第一和第三个因素相互作用影响的假设检验。第六行表示 p 的第六个值对应第二和第三个因素相互作用影响的假设检验。

3 个因素的方差分析表如图 6-6 所示。由 p 值可知：它的第一、第二、第三和第五个元素值接近于零，这说明 3 个因素及第一与第三个因素的相互作用对汽车里程的影响较显著；它的第四和第六个元素值大于零，这说明第一个与第二个因素的相互作用，以及第二与第三个因素的相互作用对汽车里程的影响不太显著。

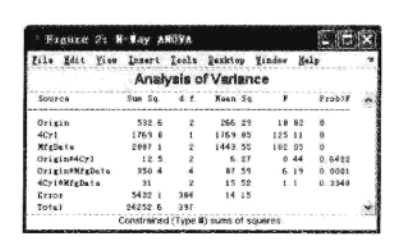


Figure 2: N-Way ANOVA

File Edit View Insert Tools Desktop Window Help

Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
Origin	532.6	2	266.29	18.82	0
4Cyl	1769.8	1	1769.85	125.11	0
MfgDate	2887.1	2	1443.55	102.05	0
Origin*4Cyl	12.5	2	6.27	0.44	0.6422
Origin*MfgDate	350.4	4	87.59	6.19	0.0001
4Cyl*MfgDate	31	2	15.52	1.1	0.3348
Error	5432.1	384	14.15		
Total	24252.6	397			

Constrained (Type III) sums of squares.

图 6-6　3 个因素的方差分析表

6.4　数据曲线拟合

6.4.1　多项式拟合

一般多项式拟合的目标是找出一组多项式系数 $a_i\ (i=1,2,\cdots,n+1)$，使得多项式

$$\psi(x)=a_1x^n+a_2x^{n-1}+\cdots+a_nx+a_{n+1} \tag{6-17}$$

能够较好地拟合原始数据。多项式拟合并不能保证每个样本点都在拟合的曲线上，但能使得整体的拟合误差较小。多项式拟合可以通过 MATLAB 提供的 polyfit 函数实现。

该函数的调用格式如下：

```
p=polyfit(x, y, n)
```

其中，x 和 y 为原始的样本点构成的向量，n 为选定的多项式的阶次，p 为多项式系数按降幂排列得出的行向量，可以用符号运算工具箱中的 poly2sym 函数将其转换成真正的多项式形式，也可以使用 polyval 函数求取多项式的值。下面通过例子演示多项式拟合函数的使用方法和优缺点。

【例 6-11】 假设已知的数据点来自函数 $f(x)=(x^2+3x+5)e^{-5x}\sin x$，试观察多项式拟合的效果。

其实现的 MATLAB 程序代码如下：

```
>> clear all;
x0=-1+2*[0:10]/10;
y0=1./(1+25*x0.^2);
x=-1:0.01:1;
ya=1./(1+25*x.^2);
p3=polyfit(x0,y0,3);y1=polyval(p3,x);
p5=polyfit(x0,y0,5);y2=polyval(p5,x);
p8=polyfit(x0,y0,8);y3=polyval(p8,x);
p10=polyfit(x0,y0,10);y4=polyval(p10,x);
plot(x,ya,x,y1,x,y2,'-.',x,y3,'--',x,y4,':');
```

运行程序，效果如图 6-7 所示。

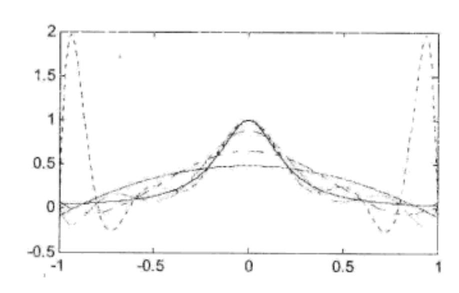


图 6-7 各阶多项式拟合的效果

该例子如果用 Taylor 幂级数展开，效果将更差。用下面的语句可以得出 Taylor 幂级数展开式及拟合效果，并可以绘制出该多项式的效果，如图 6-8 所示。可以看出，这样拟合的结果是相当差的，甚至可以说是完全错误的。

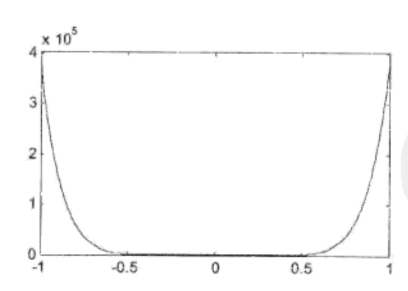


图 6-8 Taylor 幂级数展开

其实现的 MATLAB 程序代码如下：

```
>> syms x;
y=1/(1+25*x^2);
p=taylor(y,x,10)
x1=-1:0.01:1;ya=1./(1+25*x1.^2);
y1=subs(p,x,x1);plot(x1,y1);
```

运行程序，输出如下：

```
p =
1-25*x^2+625*x^4-15625*x^6+390625*x^8
```



6.4.2 连分式展开及连分式的有理近似

连分式是对函数或数值的一种很有效的近似形式。函数 $f(x)$ 经常可以用连分式表示为

$$f(x)=\cfrac{1}{f_1(x)+\cfrac{1}{f_2(x)+\cfrac{1}{f_3(x)+\cfrac{1}{f_4(x)+\cfrac{1}{\cdots}}}}} \tag{6-18}$$

最常见的连分式形式为 CauerII 型连分式，表示为

$$f(x)=\cfrac{\alpha_1}{\beta_1+\cfrac{\alpha_2 x}{\beta_2+\cfrac{\alpha_3 x}{\beta_3+\cfrac{\alpha_4 x}{\beta_4+\cfrac{\alpha_5 x}{\cdots}}}}} \tag{6-19}$$

MATLAB 语言及符号运算工具箱并未直接提供连分式展开的函数，但可以调用 Maple 中的 cfrac 函数来求取函数的连分式展开。在调用该函数前，还需要将 Maple 的数用 with 函数调入，这样给定函数或数值的 CauerII 型连分式展开可以用下面的命令实现。

```
maple('with(numtheory):')          %调入数论包
f=maple(['cfe:=cfrac(' fun ',x,m)']); %调用连分式函数,生成 cfe 变量
```

其中，该函数将 MATLAB 定义的函数字符 fun 进行连分式展开，自变量为 x，展开的项数为 m，该函数得出的部分分式展开 cfe 为 Maple 环境中的变量，而 f 为返回到 MATLAB 环境中的字符串。若对数值进行连分式展开，则可以不给出 x 变量。

由保留的前 n 级连分式的项，可以调用 Maple 中的 nthnumer 函数和 nthdenom 函数变换出有理函数的近似形式。

这两个函数的调用格式如下：

```
p=maple('nthnumer','cfe',n);   %由 cfe 变量提取前 n 级的分子
q=maple('nthdenom','cfe',n);   %由 cfe 变量提取前 n 级的分子
```

由上面两个命令，可以得出有理函数近似形式的分子和分母。

【例 6-12】 先观察一个常数的连分式近似问题，试对 π 进行 20 级近似，并找出一个较好的连分式近似阶次。

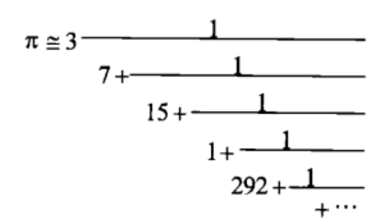
一个常数的连分式可以用下面的语句直接得出：

```
>> maple('with(numtheory):');
f=maple(['cfe:=cfrac(pi,20)'])
```

运行程序，输出如下：

```
f =
cfe:= 3+1/(7+1/(15+1/(1+1/(292+1/(1+1/(1+1/(1+1/(2+1/(1+1/(3+1/(1+1/(14+1/(2+1/
(1+1/(1+1/(2+1/(2+1/(2+1/(2+1/(1+`...`)))))))))))))))))))))
```

亦即 π 的连分式展开式子为



其中，292 和其他值相差较大，所以截断到此级即可以得出较高的精度。由有理近似的函数，则可以得出分子和分母的值为

```
>>n=maple('nthnumer','cfe',4);
  d=maple('nthdenom','cfe',4);
   [vpa(n),vpa(d)]
```

运行程序，输出如下：

```
ans =
    [ 103993.,   33102.]
```

这时，还可以得出 4 级连分式有理近似为 $\frac{103993}{33102} \approx 3.14159265301190260407226149477 37$。可见，只用 4 级连分式近似就相当接近 π 值了。

【例 6-13】 根据要求，可以用下面的语句立即得出前 10 级连分式表达式。

其实现的 MATLAB 程序代码如下：

```
>> syms x;
fun='sin(x)*exp(-x)/(x+1)^3';   %fun 应该为字符串
maple('with(numtheory):');
f=maple(['cfe:=cfrac(' fun ',x,10)'])
```

运行程序，输出如下：

```
f =
cfe:= x/(1+4*x/(1-5*x/(3+43*x/(20-337*x/(43+28274*x/(1685-66157779*x/
      (395836-9881300005*x/(512851+140501598188444*x/(158335371-
      531240292464601408*x/(2484643103+`...`))))))))))
```

亦即其展开式为



$$f(x)=\cfrac{x}{1+\cfrac{4x}{1-\cfrac{5x}{3+\cfrac{43x}{20-\cfrac{337x}{43+\cfrac{28274x}{1685-\cfrac{66157779x}{395836-\cfrac{9881300005x}{512851+\cfrac{140501598188444x}{158335371-\cfrac{531240292464601408x}{2484643103+\cdots}}}}}}}}}}$$

由下面的语句可以得出前 8 级和前 10 级分式的有理多项式近似。

```
>> n=collect(maple('nthnumer','cfe',8),x);              %分子多项式合并同类项
d=collect(maple('nthdenom','cfe',8),x);
[n,d]=numden(n/d);
G=n/d;latex(G)
n=collect(maple('nthnumer','cfe',10),x);                %分子多项式合并同类项
d=collect(maple('nthdenom','cfe',10),x);
[n,d]=numden(n/d);
G1=n/d;latex(G1)
```

显示如下：

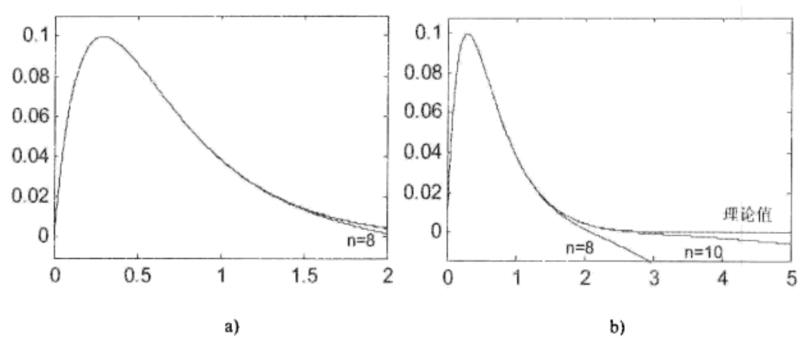
```
ans =
  10\,{\frac{x\left(845713\,{x}^{3}-4973560\,{x}^{2}+11841438\,x-10769871\right) }{5864273\,
  {x}^{4}-83147900\,{x}^{3}-294069480\,{x}^{2}-312380460\,x- 107698710}}
ans =
  -{\frac{x\left( 170455846739\,{x}^{4}-472453225650\,{x}^{3}-3615529382220\,{x}^{2}+
  20275122684600\,x-28175852788020\right) }{2071713977216\,{x}^{5}+14187032489655\,
  {x}^{4}+58214153847990\,{x}^{3}+110354057230620\,{x}^{2}+92428288467480\,x+28175852788020}}
```

这时可以得出

$$f_8(x)\cong 10\frac{x(845713x^3-4973560x^2+11841438x-10769871)}{5846273x^4-83147900x^3-294069480x^2-312380460x-107698710}$$

用下面的语句还可以得出（0, 2）区间内的原始函数 $f(x)$ 和 $f_8(x)$ 的曲线，如图 6-9a 所示。可见，拟合效果还是很理想的，n=10 时效果更好些，几乎无法区分原函数曲线和拟合曲线。若扩大拟合区域，令其为（0, 5），则可以得出如图 6-9b 所示的拟合曲线，可见这样的拟合效果变差，需要进一步增加连分式级数，所以这样的方法有时不适合于大区域拟合。

```
>> ezplot(fun,[0,2]);
hold on;
ezplot(G,[0,2]);ezplot(G1,[0,2]);
figure;ezplot(fun,[0,5]);
hold on
;ezplot(G,[0,5]);ezplot(G1,[0,5]);
```



a) b)

图 6-9 连分式拟合效果比较

a) (0, 2)区间拟合效果 b) (0, 5)区间拟合效果

6.4.3 有理式拟合

假设某函数 $f(s)$ 的幂级数展开可以表示为

$$f(s)=c_0+c_1s+c_2s^2+\cdots=\sum_{i=0}^{\infty}c_is^i \tag{6-20}$$

并假设 r/m 的有理拟合近似可以写成如下的有理函数形式

$$G_m^r(s)=\frac{\beta_{r+1}s^r+\beta_rs^{r-1}+\cdots+\beta_1}{\alpha_{m+1}s^m+\alpha_ms^{m-1}+\cdots+\alpha_1}=\frac{\sum_{i=1}^{r+1}\beta_is^{i-1}}{\sum_{i=1}^{m+1}\alpha_is^{i-1}} \tag{6-21}$$

式中，$\alpha_1=1$；$\beta_1=c_1$。设 $\sum_{i=0}^{\infty}c_is^i=G_m^r(s)$，则可以写出如下的等式

$$\sum_{i=1}^{m+1}\alpha_is^{i-1}\sum_{i=0}^{\infty}c_is^i=\sum_{i=1}^{r+1}\beta_is^{i-1} \tag{6-22}$$

对比等式中 s 相应次数的系数，令相应的 s 项系数的值相等，则 α_i ($i=1,2,\cdots,m+1$)和 β_i ($i=1,2,\cdots,r+1$)可通过下面的方程求解出来。

$$\boldsymbol{Wx}=\boldsymbol{w},\quad \boldsymbol{v}=\boldsymbol{Vy} \tag{6-23}$$

其中，

$$\begin{aligned}&\boldsymbol{x}=(\alpha_1,\ \alpha_3,\ \cdots,\ \alpha_{m+1})^{\mathrm{T}},\quad \boldsymbol{w}=(-c_{r+2},\ -c_{r+3},\ \cdots,\ -c_{m+r+1})^{\mathrm{T}}\\&\boldsymbol{v}=(\beta_2-c_2,\ \beta_3-c_3,\ \cdots,\ \beta_{r+1}-c_{r+1})^{\mathrm{T}},\quad \boldsymbol{y}=(\alpha_1,\ \alpha_3,\ \cdots,\ \alpha_{r+1})^{\mathrm{T}}\end{aligned} \tag{6-24}$$

且

$$\boldsymbol{W}=\begin{pmatrix}c_{r+1} & c_r & \cdots & 0 & \cdots & 0\\ c_{r+2} & c_{r+1} & \cdots & c_1 & \cdots & 0\\ \vdots & \vdots & & & & \vdots\\ c_{r+m} & c_{r+m-1} & \cdots & c_{m-1} & \cdots & c_{r+1}\end{pmatrix} \tag{6-25}$$



$$V=\begin{pmatrix} c_1 & 0 & 0 & \cdots & 0 \\ c_2 & c_1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ c_r & c_{r-1} & c_{r-2} & \cdots & c_1 \end{pmatrix} \tag{6-26}$$

可以证明，若有理拟合近似的分子分母阶次相同或分母比分子高一阶，则该近似等效于GauerII 型连分式近似。可以通过 MATLAB 的 padefcn 函数计算 $f(x)$ 的有理式拟合函数近似。该函数的 MATLAB 程序代码如下：

```
function [nP,dP]=padefcn(c,r,m)
w=-c(r+2:m+r+1)';
vv=[c(r+1:-1:1)';zeros(m-1-r,1)];
W=rot90(hankel(c(m+r:-1:r+1),vv));
V=rot90(hankel(c(r:-1:1)));
x=[1 (W\w)'];
y=[1 x(2:r+1)*V'+c(2:r+1)];
dP=x(m+1:-1:1)/x(m+1);nP=y(r+1:-1:1)/x(m+1);
```

【例 6-14】 试对 $f(x)=\mathrm{e}^{-2x}$ 函数用有理式拟合函数近似。

解： 可以选择不同的分母阶次，选择分子阶次为 0，并选择不同的分母阶次，则可以得出不同的有理式拟合近似式，近似曲线如图 6-10 所示。

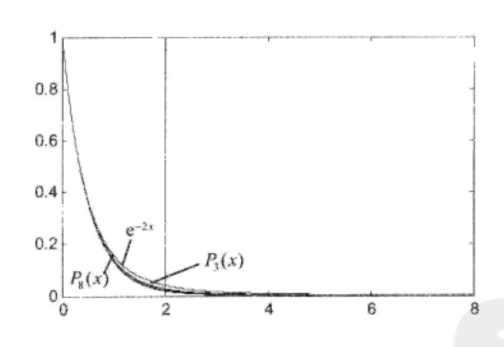


图 6-10 原始数据与拟合曲线

其实现的 MATLAB 程序代码如下：

```
>> syms x;
c=taylor(exp(-2*x),10);
c=sym2poly(c);
c=c(end:-1:1);
x=0:0.01:8;
nd=[3:7];xx=[0,2,2+eps,8];
yy=[0,0,1,1];plot(xx,yy);
hold on;
for i=1:length(nd)
    [n,d]=padefcn(c,0,nd(i));
```

```
    y=polyval(n,x)./polyval(d,x);
    plot(x,y);
end
```

由图 6-9 可见，3 阶近似得出的效果尚可，如果增加阶次，会得出更好的效果，8 阶近似的结果还是很精确的。8 阶有理式拟合近似表达式如下：

$$P_8(x)=\frac{157.5}{x^8+4x^7+14x^6+42x^5+105x^4+210x^3+315x^2+315x+157.5}$$

6.4.4 函数线性组合的曲线拟合方法

假设已知某函数的线性组合为

$$g(x)=c_1f_1(x)+c_2f_2(x)+\cdots+c_nf_n(x) \tag{6-27}$$

式中，$f_1(x),f_2(x),\cdots,f_n(x)$ 为已知函数，$c_1,c_2,\cdots,c_n$ 为待定系数。

这时假设已经测出数据 $(x_1y_1,x_2y_2,\cdots,x_My_M)$，则可以建立如下的线性方程

$$\boldsymbol{Ac}=\boldsymbol{y} \tag{6-28}$$

式中，

$$\boldsymbol{A}=\begin{pmatrix} f_1(x_1) & f_2(x_1) & \cdots & f_m(x_1) \\ f_1(x_2) & f_2(x_2) & \cdots & f_m(x_2) \\ \vdots & \vdots & & \vdots \\ f_1(x_M) & f_2(x_M) & \cdots & f_m(x_M) \end{pmatrix},\quad \boldsymbol{y}=\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_M \end{pmatrix} \tag{6-29}$$

且 $\boldsymbol{c}=(c_1,\ c_2,\ \cdots,\ c_n)^{\mathrm{T}}$。故该方程的最小二乘解为 $\boldsymbol{c}=\dfrac{\boldsymbol{y}}{\boldsymbol{A}}$。

【例 6-15】 假设测出了一组 x_i，y_i，由下表给出，且已知函数原型 $y(x)=c_1+c_2\mathrm{e}^{-3x}+c_3\cos(-2x)\mathrm{e}^{-4x}+c_4x^2$，试用已知数据求出待定系数 c_i 的值。

x_i	0	0.2	0.4	0.7	0.9	0.92	0.99	1.2	1.4	1.48	1.5
y_i	2.88	2.2576	1.9683	1.9258	2.0862	2.109	2.1979	2.5409	2.9627	3.155	3.2052

其实现的 MATLAB 程序代码如下：

```
>> x=[0,0.2,0.4,0.7,0.9,0.92,0.99,1.2,1.4,1.48,1.5]';
y=[2.88,2.2576,1.9683,1.9258,2.0862,2.109,2.1979,2.5409,2.9627,3.155,3.2052]';
A=[ones(size(x)),exp(-3*x),cos(-2*x).*exp(-4*x),x.^2];
c=A\y;
c1=c'
```

运行程序，输出如下：

```
c1 =
    1.2200    2.3397    -0.6797    0.8700
>> x0=[0:0.01:1.5]';
A1=[ones(size(x0)),exp(-3*x0),cos(-2*x0).*exp(-4*x0),x0.^2];
y1=A1*c;
```

```
plot(x0,y1,x,y,'x');
```

这时可以得出拟合曲线和已知数据点，如图 6-11 所示。可见拟合效果是令人满意的。

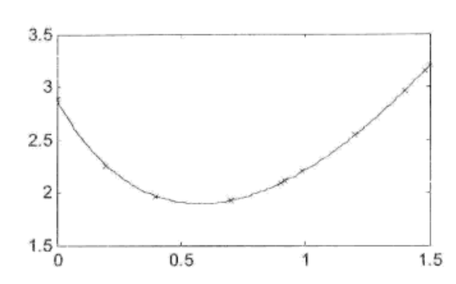


图 6-11 原始数据与拟合曲线

【例 6-16】 假设测出一组实际数据，试对其进行函数拟合。

x_i	1.1052	1.3499	1.4918	1.6487	1.8221	2.0138	2.2255	2.4596	2.7183	3.6693
y_i	0.6795	0.5309	0.4693	0.4148	0.3666	0.3241	0.2865	0.2532	0.2238	0.1546

解：可以用下面的语句将表中给出的数据用曲线表示出来，如图 6-12a 所示。

其实现的 MATLAB 程序代码如下：

```
>> x=[1.1052  1.3499  1.4918  1.6487  1.8221  2.0138 2.2255 2.4596 2.7183 3.6693];
y=[0.6795  0.5309 0.4693 0.4148 0.3666 0.3241  0.2865  0.2532  0.2238  0.1546];
plot(x,y,x,y,'*');
```

在实际曲线拟合时，有时从 x, y 本身看不出它们之间的关系，则可能需要对数据进行可能的非线性变换，观察是否得出线性关系。例如，可以对 x, y 分别进行对数变换，得出如图 6-12b 所示的曲线，可见二者是线性的。

```
>> x1=log(x);y1=log(y);
plot(x1,y1,x1,y1,'*')
```

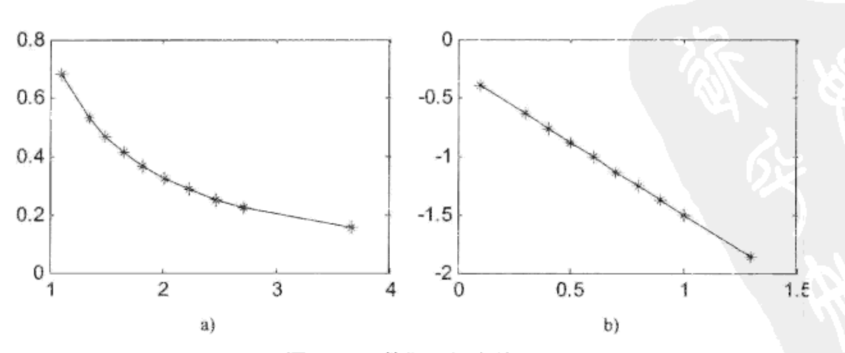




图 6-12 数据及拟合结果

a) 曲线拟合 b) 对数变换后的拟合

其中，两个空矩阵表示 α 向量的上下限。由于对这些参数的范围无限制，故采用了默认

的表示形式。可以看出，修改误差限后，得出的拟合待定系数更加精确。绘制出的拟合曲线与样本点如图 6-13 所示。

```
>> A=[x1',ones(size(x1'))];
   c=[A\y1']
c =
    -1.2339     -0.2630
>> exp(c(2))
ans =       0.7687
```

【例 6-17】 多项式拟合可以认为是前面介绍的多函数线性组合的特例，这样可以选择各个函数为 $f_i(x)=x^{n+1-i}$ ($i=1,2,\cdots,n$)，用该方法重新考虑例 6-11 中数据的多项式拟合问题，试观察多项式拟合的效果。

其实现的 MATLAB 程序代码如下：

```
>> x=[0:0.1:2]';
y=(x.^2-3*x+5).*exp(-5*x).*sin(x);
n=7;A=[];
for i=1:n+1
      A(:,i)=x.^(n+1-i);
end
c=A\y;vpa(poly2sym(c),5)
```

运行程序，输出如下：

```
ans =
   .90419*x^7-7.2884*x^6+24.001*x^5-41.422*x^4+39.735*x^3-20.298*x^2+
   4.3877*x+.35535e-2
```

6.4.5 最小二乘曲线拟合

假设有一组数据 x_i, y_i ($i=1,2,\cdots,N$)，且已知这组数据满足某一函数原型 $\hat{y}(x)=f(a,x)$，其中 a 为待定系数向量，则最小二乘曲线拟合的目标就是求出这一组待定系数的值，使得目标函数

$$J=\min_a \sum_{i=1}^{N}\left[y_i-\hat{y}(x_i)\right]^2=\min_a \sum_{i=1}^{N}\left[y_i-f(a,x_i)\right]^2 \tag{6-30}$$

最小。MATLAB 的统计工具箱提供了 lsqcurvefit 函数，可以解决最小二乘曲线拟合的问题。

该函数的调用格式如下：

```
[a, Jm]=lsqcurvefit(Fun, a0, x, y)
```

其中，Fun 为原型函数的 MATLAB 表示，可以是 M-函数或 inline 函数；$a0$ 为最优化的初值；x, y 为原始输入输出数据向量。调用该函数，将返回待定系数向量 a，以及在此待定系数下的目标函数的值 Jm。

【例 6-18】 假设由下面的语句生成一组数据 x 和 y。

```
>> x=0:0.1:10;
y=0.12*exp(-0.213*x)+0.54*exp(-0.17*x).*sin(1.23*x);
```

并已知该数据满足的原型函数为 $y(x)=a_1e^{-a_2x}+a_3e^{-a_4x}\sin(a_5x)$，其中，$a_i$ 为待定系数。采用最小二乘曲线拟合的目的就是获得这些待定系数，使得目标函数的值为最小。



根据已知的函数原型，可以编写出如下的 MATLAB 程序代码：

```
>>  x=0:0.1:10;
y=0.12*exp(-0.213*x)+0.54*exp(-0.17*x).*sin(1.23*x);
f=inline('a(1)*exp(-a(2)*x)+a(3)*exp(-a(4)*x).*sin(a(5)*x)','a','x');
%建立起函数的原型,则可以由下面的语句得出待定系数向量了
[xx,res]=lsqcurvefit(f,[1,1,1,1,1],x,y);
xx',res
```

运行程序，输出如下：

```
Optimization terminated: first-order optimality less than OPTIONS.TolFun,
 and no negative/zero curvature detected in trust region model.
ans =
    0.1200
    0.2130
    0.5400
    0.1700
    1.2300
res =   1.7928e-016
```

可以看出，这样得出的待定系数精度较高，接近于理论值 $a=(0.12,0.213,0.54,0.17,1.23)^{\mathrm{T}}$。如果想进一步提高精度，则需要修改最优化的选项，这时函数的调用格式也将发生变化。

```
>>  %修改精度限制
ff=optimset;ff.TolFun=1e-20;
ff.TolX=1e-15;
[xx,res]=lsqcurvefit(f,[1,1,1,1,1],x,y,[],[],ff);
xx',res
```

运行程序，输出如下：

```
Optimization terminated: first-order optimality less than OPTIONS.TolFun,
 and no negative/zero curvature detected in trust region model.
ans =
    0.1200
    0.2130
    0.5400
    0.1700
    1.2300
res =       0
>> x1=0:0.01:10;y1=f(xx,x1);
plot(x1,y1,x,y,'o')
```

其中，两个空矩阵表示 a 向量的上下限。由于对这些参数的范围无限制，故采用了默认的表示形式。可以看出，修改误差限后，得出的拟合待定系数更加精确。绘制出的拟合曲线与样本点如图 6-13 所示。

【例 6-19】 假设有一组实测数据，如下所示：

x_i	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
y_i	2.3201	2.6470	2.9707	3.2885	3.6008	3.9090	4.2147	4.5191	4.8232	5.1275

假设已知该数据可能满足的原型函数为 $y(x)=ax+bx^2\mathrm{e}^{-cx}+d$，试求出满足下面数据的最小二乘解 a,b,c,d 的值。

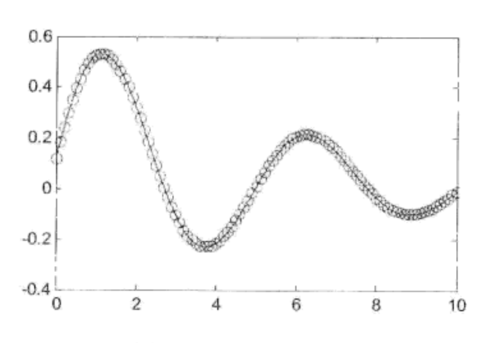


图 6-13　拟合效果比较

其实现的 MATLAB 程序代码如下：

```
>> x=0.1:0.1:1;
y=[2.3201 2.6470 2.9707    3.2885 3.6008    3.9090    4.2147    4.5191    4.8232    5.1275];
```

令 $a_1=a,a_2=b,a_3=c,a_4=d$，这样，原型函数可以写成 $y(x)=a_1x+a_2x^2e^{-a_3x}+a_4$，可以用 MATLAB 程序代码写出：

```
function y=c8f3(a,x)
y=a(1)*x+a(2)*x.^2.*exp(-a(3)*x)+a(4);
```

则

```
>> a=lsqcurvefit('c8f3',[1;2;2;3],x,y);
>> a'
Optimization terminated: relative function value
 changing by less than OPTIONS.TolFun.
ans =
    3.1001    1.5027    4.0046    2.0000
```

用下面的语句还可以计算出各个点处的值，可以将两曲线绘制在同一坐标系下，如图 6-14 所示。可见，两曲线还是很接近的，说明拟合效果较好。

```
>> y1=c8f3(a,x);plot(x,y,x,y1,'o')
```



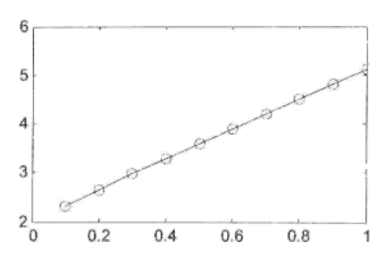


图 6-14　拟合效果比较



6.5　二次响应曲面模型

响应曲面方法是定量表示多个输入变量与一个输出变量之间关系的一种有效工具。假设一个输出 z 是两个输入 x,y 的多项式函数，那么函数 $z=f(x,y)$ 是空间 (x,y,z) 的一个二维曲面。

对 3 个输入 x_1,x_2,x_3 来说，二次响应曲面的方程为：

$$y=b_0+b_1x_1+b_2x_2+b_3x_3+\cdots+ \qquad \text{（线性项）}$$

$$b_{12}x_1x_2+b_{13}x_1x_3+b_{23}x_2x_3+\cdots+ \qquad \text{（交叉项）}$$

$$b_{11}x_1^2+b_{22}x_2^2+b_{33}x_3^2 \qquad \text{（二次项）}$$

统计工具箱提供了用于交互式拟合和显示响应曲面的 rstool 函数。

其调用格式如下：

```
rstool(x,y)
rstool(x, y, model)
rstool(x, y, model, alpha, 'xname', 'yname')
```

其中，x,y 是输入数据；model 是模型的种类，其取值如下：

- model='linear'：表示仅仅包括常数项和一次项。
- model='purequadratic'：表示包括常数项、一次项和二次项。
- model='interaction'：表示包括常数项、一次项和交叉项。
- model='quadratic'：表示包括交叉项和二次项。
- alpha 是置信水平；'xname'是 x 轴的标记；'yname'是 y 轴的标记。

下面通过一个例子，说明 rstool 函数的用法。文件 reaction.mat 中包含的数据反映的是某化学过程，它是 3 个化学反应物（氢、戊烷和异戊烷）压力的函数。利用 rstool 函数可以分别显示这 3 个压力与反应率之间的关系曲线。

```
>> %装载数据
load reaction
%设置参数
model='quadratic';
alpha0=0.01;
%显示
```

```
rstool(reactants,rate,model,alpha0,xn,yn);
```

利用 4 种不同的模型，拟合得到的关系曲线分别如图 6-15~图 6-18 所示。在每个图中分别有 3 幅子图，对应了 3 个变量与反应率之间的关系曲线。在每幅子图中，其他两个变量的值固定，且在下面的可编辑的文本框中显示。任意改变其他两个变量的值，对应的子图会刷新，显示新的图形。

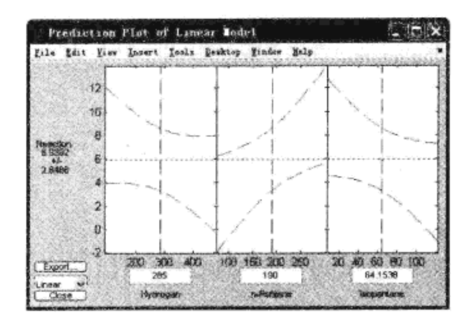


图 6-15　model='linear'时的关系曲线

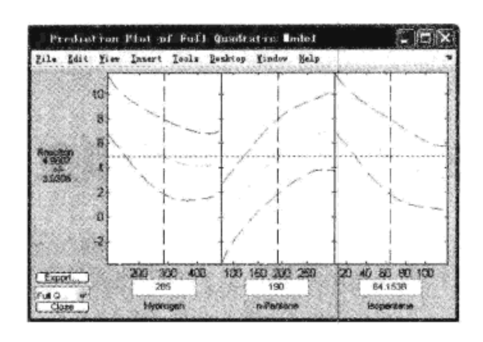


图 6-16　model='quadratic'时的关系曲线

另外，还可以通过单击“Export”按钮，将计算得到的变量保存到工作空间，拟合得到的系数按照如下的顺序：

1）常数项。2）线性项。3）交叉项。4）二次项。

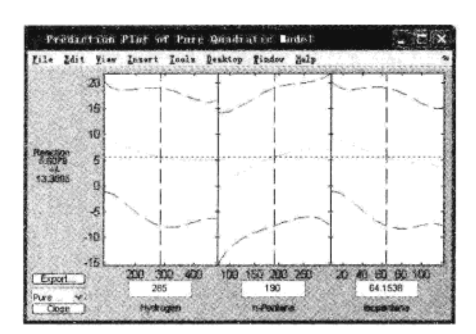


图 6-17　model='purequadratic'时的关系曲线

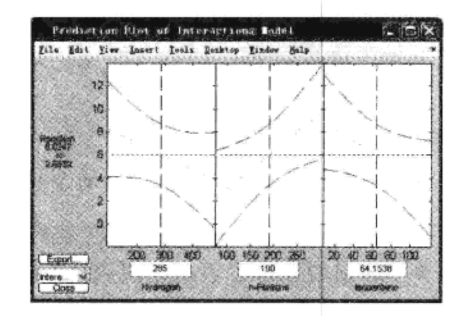


图 6-18　model='interaction'时的关系曲线

第7章 回归分析



7.1 一元线性回归分析

在许多实际中，经常需要寻找两个（或多个）变量之间的关系，并希望利用观测数据拟合系统的数学模型，其中最简单的模型是线性模型。

7.1.1 一元线性回归分析的基本定义

假设随机变量 y 和 x 之间服从以下的线性关系

$$y = \alpha + \beta x + \varepsilon \tag{7-1}$$

现存在 n 个值 $y_i, x_i (i = 1, 2, \cdots, n)$，则它们满足关系

$$y_i = \alpha + \beta x_i + \varepsilon \tag{7-2}$$

假设 ε_i 相互独立且满足

$$\varepsilon_i \sim N(0, \sigma^2)，\quad i = 1, 2, \cdots, n \tag{7-3}$$

则称变量 y 和 x 服从一元线性回归模型（或一元线性正态回归模型）。

对上述定义的一元线性回归模型，实际考虑的统计推断问题是：在已知观测值 $y_i, x_i (i = 1, 2, \cdots, n)$ 的基础上，对未知参数 α，β，σ^2 进行估计，对 α，β 的某种假设进行检验，对 y 进行预报等。

7.1.2 未知参数估计

（1）(α, β) 的最小二乘估计

对一组观测值 $y_i, x_i (i = 1, 2, \cdots, n)$，它满足

$$y_i = \alpha + \beta x_i + \varepsilon_i \tag{7-4}$$

最小二乘法是寻找未知参数 (α, β) 的估计量 $(\hat{\alpha}, \hat{\beta})$，使得

$$\sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta} x_i)^2 = \min_{\alpha, \beta} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 \tag{7-5}$$

满足式（7-5）的估计量 $(\hat{\alpha}, \hat{\beta})$ 被称为 (α, β) 的最小二乘估计。

记

$$P(\alpha, \beta) = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 \tag{7-6}$$



令

$$\frac{\partial P}{\partial \alpha}=0\text{，}\quad \frac{\partial P}{\partial \beta}=0 \tag{7-7}$$

可以得到

$$\begin{cases} n\hat{\alpha}+n\overline{x}\hat{\beta}=n\overline{y} \\ n\overline{x}\hat{\alpha}+\sum_{i=1}^{n}x_i^2\hat{\beta}=\sum_{i=1}^{n}x_iy_i \end{cases} \tag{7-8}$$

式中，$\overline{x}=\frac{\sum_{i=1}^{n}x_i}{n}$；$\overline{y}=\frac{\sum_{i=1}^{n}y_i}{n}$。

为简化记号，令

$$l_{xy}=\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})=\sum_{i=1}^{n}x_iy_i-\frac{1}{n}\sum_{i=1}^{n}x_i\cdot\sum_{i=1}^{n}y_i \tag{7-9}$$

$$l_{xx}=\sum_{i=1}^{n}(x_i-\overline{x})^2=\sum_{i=1}^{n}x_i^2-\frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2 \tag{7-10}$$

求解方程，得到唯一解为

$$\begin{cases} \hat{\alpha}=\overline{y}-\hat{\beta}\overline{x} \\ \hat{\beta}=\dfrac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2} \end{cases} \tag{7-11}$$

在平面直角坐标系中，通过（0, $\hat{\alpha}$）与（$\overline{x},\overline{y}$）两点引一直线，即为所求的回归直线。这是因为点（0, $\hat{\alpha}$）显然在直线

$$\hat{y}=\hat{\alpha}+\hat{\beta}x$$

上。若将 $\hat{\alpha}=\hat{y}-\hat{\beta}x$ 代入式（7-11），则有

$$\hat{y}-\overline{y}=\hat{\beta}(x-\overline{x}) \tag{7-12}$$

可知点（$\overline{x},\overline{y}$）也在这条直线上。

（2）(α,β) 的最小二乘估计的矩阵算法

一元线性回归模型参数的最小二乘估计的矩阵算法记为

$$\boldsymbol{y}=\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},\quad \boldsymbol{X}=\begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix},\quad \boldsymbol{A}=\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} \tag{7-13}$$

则一元线性回归的数据模型为 $\boldsymbol{y}=\boldsymbol{X}\boldsymbol{A}$ 。这是一个不相容的线性方程组，当 rank($\boldsymbol{X}$)=2<n 时，其最小二乘解为

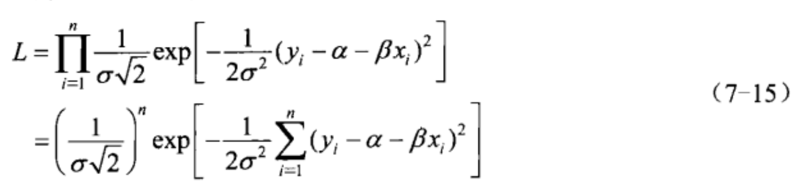
$$\boldsymbol{A}=(\boldsymbol{X}^{\mathrm{T}}-\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}-\boldsymbol{y} \tag{7-14}$$

通常，在高等代数的广义逆矩阵理论中有关于这一算法的详细推证。感兴趣的读者请自行查阅相关资料。

（3）(α,β) 的极大似然估计



由于 y_i 相互独立，且 $y_i \sim N(\alpha+\beta x_i,\sigma^2)$，则有 $y_1,y_2,\cdots,y_n$ 的联合概率密度为



$$\begin{aligned} L &= \prod_{i=1}^{n}\frac{1}{\sigma\sqrt{2}}\exp\left[-\frac{1}{2\sigma^2}(y_i-\alpha-\beta x_i)^2\right] \\ &= \left(\frac{1}{\sigma\sqrt{2}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\alpha-\beta x_i)^2\right] \end{aligned} \tag{7-15}$$

要求估计的 $(\hat{\alpha},\hat{\beta})$ 使得似然函数 L 取得最大值，只要

$$P(\alpha,\beta)=\sum_{i=1}^{n}(y_i-\alpha-\beta x_i)^2 \tag{7-16}$$

取得最小值即可。这又回到了最小二乘估计的情形。

（4）σ^2 的估计

由于 $\sigma^2=D\varepsilon=E\varepsilon^2$，故可以用 $\dfrac{\sum_{i=1}^{n}\varepsilon_i^2}{n}$ 对 σ^2 作矩估计，以 α,β 的相应估计量代入，可得

$$\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{\alpha}-\hat{\beta}x_i)^2 \tag{7-17}$$

式（7-17）可以看做是近似矩估计。

代入 $(\hat{\alpha},\hat{\beta})$ 的估计值，则有

$$\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^2-\beta\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2\right) \tag{7-18}$$

7.1.3 回归方程的显著性检验

建立经验回归方程的目的在于揭示两个相关变量 x 与 y 之间的内在规律，然而，对任意样本观测值 $x_i,y_i\ (i=1,2,\cdots,n)$ 做出的散点图，即使一看就知道 x 与 y 之间根本不存在线性关系，也能由式（7-11）算出 $\hat{\alpha}$，$\hat{\beta}$，从而写出线性回归方程 $\hat{y}=\hat{\alpha}+\hat{\beta}x$，但这时所建立的回归方程是毫无意义的。什么是一个有意义的回归方程呢？首先注意到 $y=\alpha+\beta x+\varepsilon$，当 $|\beta|$ 越大，y 随 x 的变化越显著；当 $|\beta|$ 越小，y 随 x 的变化越不明显。特别当 $\beta=0$ 时，意味着 y 与 x 之间没有线性关系。也就是说，所建立的回归方程没有意义；因此当 $\beta\neq 0$ 时，所建立的回归方程才有意义。这实质上就是要对假设 H_0： $\beta=0$ 进行检验，这种检验称为回归显著性检验。

为了寻找合适的统计量，对关系式 $l_{yy}=\sum_{i=1}^{n}(y_i-\overline{y})^2$ 进行分解，并称 l_{yy} 为总的偏差平方和，记作 S_T，它反映 $y_1,y_2,\cdots,y_n$ 的离散程度，即

$$S_T = \sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2 \tag{7-19}$$

由于变量 y 的各个观测值 y_i 与其均值 $\overline{y}$ 的离差 $y_i - \overline{y}$ 可以分解为两部分

$$y_i - \overline{y} = (\hat{y}_i - \overline{y}) + (y_i - \hat{y}_i)$$

式中，$y_i - \overline{y}$ 是 y_i 与 $\overline{y}$ 的离差；$\hat{y}_i - \overline{y}$ 是回归值 $\hat{y}_i$ 与均值 $\overline{y}$ 的离差，这是回归能解释的部分；$y_i - \hat{y}_i$ 是观测值 y_i 与回归值 $\hat{y}_i$ 的离差，亦即残差 e_i，这是回归不能解释的部分。

因为

$$\sum_i (y_i - \overline{y})^2 = \sum_i (\hat{y}_i - \overline{y})^2 + \sum_i (y_i - \hat{y}_i)^2 + 2\sum_i (\hat{y}_i - \overline{y})(y_i - \hat{y}_i)$$

能够证明 $\sum_i (\hat{y}_i - \overline{y})(y_i - \hat{y}_i) = 0$，因此有

$$S_T = l_{yy} = \sum_i (y_i - \overline{y})^2 = \sum_i (\hat{y}_i - \overline{y})^2 + \sum_i (y_i - \hat{y}_i)^2$$

记

$$S_R = \sum_i (\hat{y}_i - \overline{y})^2 \tag{7-20}$$

$$S_e = \sum_i (y_i - \hat{y}_i)^2 \tag{7-21}$$

于是

$$S_T = S_R + S_e \tag{7-22}$$

可以证明 $\overline{y} = \frac{1}{n}\sum_i \hat{y}_i$，因此 $S_R = \sum_i (\hat{y}_i - \overline{y})^2$ 反映回归值 $\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_n$ 的离散程度，称为回归平方和。而 $\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_n$ 的离散性又来源于 $x_1, x_2, \cdots, x_n$ 的离散性，实际上

$$\begin{aligned} S_R &= \sum_i (\hat{y}_i - \overline{y})^2 = \sum_i [(\hat{\alpha} + \hat{\beta} x_i) - (\hat{\alpha} + \hat{\beta}\overline{x})]^2 = \sum_i \hat{\beta}^2 (x_i - \overline{x})^2 = \hat{\beta}^2 \sum_i (x_i - \overline{x})^2 \\ &= \hat{\beta}^2 l_{xx} = \hat{\beta} l_{xy} \end{aligned} \tag{7-23}$$

这里 $l_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2$ 反映了 $x_1, x_2, \cdots, x_n$ 的离散程度，从而可知 $S_R = \sum_i (\hat{y}_i - \overline{y})^2$ 实际上反映了由于 x 的变化而引起 y 的波动的大小。这里，是通过 x 对 y 的相关性而引起的。

$S_e = \sum_i (y_i - \hat{y}_i)^2$ 反映了观测值与回归值之间的偏离，且等于 $P(\alpha, \beta)$ 的最小值 $\left(P(\alpha, \beta) = \sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2\right)$；反映了除 x 对 y 的线性影响之外的剩余因素对 y 引起的波动的大小。故称 $S_e = \sum_i (y_i - \hat{y}_i)^2$ 为剩余平方和（或残差平方和）。



$$S_e = S_T - S_R \tag{7-24}$$

若回归方程有意义，即 y 的波动主要是由 x 的变化引起的，其他一切因素是次要的。即要求 S_R 尽可能大，而 S_e 尽可能小。

可以证明：

1） $\frac{S_e}{\sigma^2} \sim \chi^2(n-2)$。

2） $\beta=0$ 时，$\frac{S_R}{\sigma^2} \sim \chi^2(1)$。

3） S_R 与 S_e 相互独立。

1. *F* 检验法——方差分析法

由前面的分析可知，在 H_0： $\beta=0$ 为真时：

$$F = \frac{S_R}{S_e/(n-1)} \sim F(1, n-2) \tag{7-25}$$

当 H_0 不真时，$\frac{S_R}{S_e/(n-1)}$ 有变大趋势，因而 F 也有变大趋势，故应取单侧拒绝域。对给定的显著性水平 α，当 $F \geqslant F_\alpha(1, n-2)$ 时，认为 $\beta=0$ 不真，称方程是显著的；反之，方程不显著。这种用 F 检验对回归方程作显著性检验的方法称为方差分析。其检验过程可由一张“方差分析表”来进行，见表 7-1。

表 7-1　方差分析表

方差来源	偏差平方和	自由度	方差	F 值	F_α	显著性
回归	S_R	1	$V_R=\frac{S_R}{1}$	$F=\frac{V_R}{V_e}$	$F_{0.05}(1,n-2)$	
剩余	$S_e=S_T-S_R$	$n-2$	$V_e=\frac{S_e}{n-2}$		$F_{0.01}(1,n-2)$	
部和	S_T	$n-2$				

2. *r* 检验法——拟合程度的测定

变量 y 的各个观测值点聚在回归直线 $\hat{y}=\alpha+\hat{\beta}x$ 周围的紧密程度，称为回归直线对样本数据点的拟合程度，通常用可决系数（也称为测定系数）r^2 来表示。

显然，变量 y 的各个观测值点与回归直线越靠近，S_R 在 S_T 中所占的比重就越大，因而定义

$$r^2 = \frac{S_R}{S_T} = \frac{\sum(\hat{y}_i - \overline{y})^2}{\sum(y_i - \hat{y}_i)^2} \tag{7-26}$$

它可用来测定回归直线对各观测值点的拟合程度。若全部的观测值点 $y_i (i=1,2,\cdots,n)$ 都落在回归直线上，则剩余平方和 $S_e=0$，$r^2=1$；若 x 完全无助于解释 y 的偏差，则回归平方和 $S_R=0$，$r^2=0$。显然，r^2 越接近于 1，用 x 的变化解释 y 的偏差的部分就越多，表明回归直线和各观测值点越接近，回归直线的拟合程度越高。可决系数 r^2 在[0, 1]上取值。

回归直线对样本数据点拟合程度的另一测度是线性相关系数r。在一元线性回归中，线性相关系数r实际上是可决系数r^2的平方根，即

$$r = \pm\sqrt{r^2} \tag{7-27}$$

r的符号与回归系数$\hat{\beta}$的符号相同，$|r|$越接近于1，表明回归直线对样本数据点的拟合程度越高。

3. 估计标准差

可决系数r^2和线性相关系数r描述了回归直线对样本数据点的拟合程度，但没有表示出变量y的诸观察值y_i与回归直线$\hat{y}_i = \alpha + \hat{\beta}x_i$的绝对离差数额。定义

$$S_y^2 = \frac{\sum(y_i - \hat{y}_i)^2}{n-2} = \frac{\sum e_i^2}{n-2}$$

为最小二乘残差值e_i方差，定义

$$S_y = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{\sum e_i^2}{n-2}} \tag{7-28}$$

为变量y对x的最小二乘回归的估计标准误差，简称估计标准误差。S_y^2和S_y可以作为y值与回归直线变差的测度。S_y的计量单位与变量y的单位相同。显然，S_y越小，表明误差越小。

MATLAB 提供了线性回归模型的建模与评价函数 regress。

regress 函数可用于p个自变量、一个因变量的线性回归模型，$y = X\beta + \varepsilon$，$\varepsilon \sim N(p, \sigma^2 I)$的建模和模型评价，$p$是指$p$个自变量的$n$个观测值。

其调用格式如下：

```
[b, bint, r, rint, stats]=regress(y, x, alpha)
```

其中，输入参数x表示p个自变量的n个观测值的$n \times p$矩阵；y表示因变量的n个观测值的$n \times 1$个向量，alpha 是显著性水平（可以省略，默认值为 0.05）。输出参数b返回的是模型系数（向量）β的最小二乘估计值，bint 是β的 100(1−alpha)%的置信区间，r是模型拟合残差（向量），rint 是模型拟合残差的 100(1−alpha)%的置信区间，stats 包含可决系数r^2的值，方差分析的F统计量的值、方差分析的显著性概率p的值和模型方差σ^2的估计值。其中，bint、r、rint 和 stats 可以默认。

【例 7-1】 某种合金强度与碳含量有关，研究人员在生产试验中收集了该合金的强度y与碳含量x的数据（见表 7-2）。试建立y与x的函数关系模型，并检验模型的可信度，检查数据中有无异常点。

表 7-2 合金强度与碳含量的数据表

x	0.10	0.11	0.12	0.13	0.14	0.15	0.16	0.17	0.18	0.20	0.21	0.23
y	42.0	41.5	15.0	45.5	45.0	47.5	49.0	55.0	50.0	55.0	55.5	60.5

分析：本问题的目的是确定合金强度与碳含量之间的相关系数。现已给出一组统计观测数据，通过作数据的散点图，观察散点图的形状可知，可建立一元线性回归模型，设一元线



性回归模型为 $y=\beta_0+\beta_1 x$，调用 regress 函数求解。模型的可信度可用可决系数的大小表示，因此计算出可决系数 r^2 即可。

其实现的 MATLAB 程序代码如下：

```
>> clear all;
x1=0.1:0.01:0.18;
x2=[x1,0.2,0.21,0.23]';
y=[42.0,41.5,45.0,45.5,45.0,47.5,49.0,55.0,50.0,55.0,55.5,60.5]';
x=[ones(12,1),x2];
%作数据的散点图
figure;
plot(x2,y,'+');
%回归分析
[b,bint,r,rint,stats]=regress(y,x);
b,bint,stats,
%作残差分析图
figure(2);
rcoplot(r,rint);hold on;
%预测及作回归线图
z=b(1)+b(2)*x2;
plot(x2,y,'*',x2,z,'r');
legend('预测图','回归线图');
```

运行程序，输出如下：

```
b =
    27.0269
   140.6194
bint =
    22.3226    31.7313
   111.7842   169.4546
stats =
      0.9219   118.0670      0.0000      3.1095
```

残差图如图 7-1 所示，散点图及回归线图如图 7-2 所示。

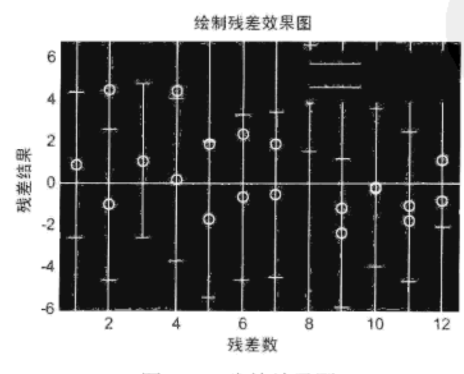


图 7-1　残差效果图

结果表明，参数的估计值 $\hat{\beta}_0=27.0269$，$\hat{\beta}_1=140.6194$；$\hat{\beta}_0$ 的置信区间为[22.3226, 31.7313]，$\hat{\beta}_1$ 的置信区间为[111.7842, 169.4546]；可决系数 $r^2=0.9219$（接近于常数 1），且 $F=118.0670$，$p=0.0000<0.05$，$\hat{\sigma}^2=3.1095$，故回归模型

$$y=27.0269+140.6194\,x$$

成立。

从图 7-1 中可看出，除第八个数据外，其余数据的残差离零点都较近，且残差的置信区间均包含零点，这说明回归模型

$$y=27.0269+140.6194\,x$$

能较好地拟合数据，而第八个数据可视为异常点。从图 7-2 中也可看出，回归线能较好地表示散点图的形状，只有第八个数据点离回归线较远。为什么会出现异常点呢？这需要对实现过程进行分析，进一步查明原因。

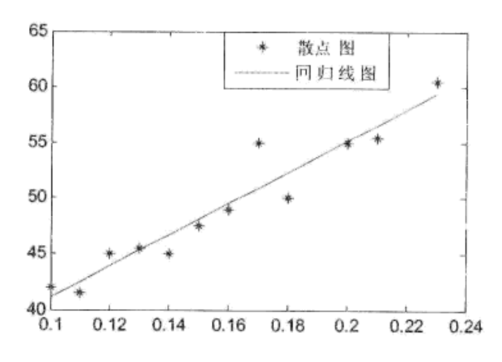


图 7-2　散点图及回归线图

7.1.4 利用回归方程进行预测

建立回归方程的目的不仅是描述变量之间的关系，更重要的是回归方程的应用。利用所建立的回归方程对因变量进行预测是其应用的基本内容。在一元线性回归分析中，当回归方程 $\hat{y}=\hat{\alpha}+\hat{\beta}x$ 具有统计显著性时，利用回归方程容易实现对因变量 y 的预测，而这一问题的实质是对 y 的点估计和区间估计。

在前面讲解的基础上，容易证明：

$$\hat{y}=\hat{\alpha}+\hat{\beta}x\sim N\left(\alpha+\beta x,\left(\frac{1}{n}+\frac{(x-\overline{x})^2}{l_{xx}}\right)\sigma^2\right)$$，且 $\hat{y}$ 与 y 相互独立。

这个结论表明，其经验回归方程 $\hat{y}=\hat{\alpha}+\hat{\beta}x$ 是线性函数 $E(y)=\alpha+\beta x$ 的无偏估计。

因此，当 $x=x_0$ 时，因变量 y 的预测值即为 $\hat{y}_0=a+bx_0$，它是 $y_0=a+bx_0+\varepsilon_0$ 的无偏估计。在显著性水平 α 下，y_0 的估计边际误差（区间估计）可由准则式

$$P\{|y_0-\hat{y}_0|<\delta\}\geqslant 1-\alpha$$

确定，由 y 和 $\hat{y}$ 的分布可以推出



$$\delta = t_{1-\frac{\alpha}{2}}(n-2)\hat{\sigma}^*\sqrt{1+\frac{1}{n}+\frac{(x_0-\overline{x})^2}{l_{xx}}}$$

显然，预测的精度取决于δ的大小，而影响δ大小的因素主要是样本容量n，x_0与x的距离及自变量的偏差平方和l_{xx}。当样本容量n较大，x_0与x的距离较近，自变量的偏差平方和l_{xx}较大（采样较为分散）时，δ的取值就较小，此时预测的精度较高。另外，当$x_0 \notin [x_{(1)}, x_{(n)}]$时，预测精度可能变得很差，在这种情况下需要特别小心。

由于上面的计算边际误差δ的公式冗繁，故在实际应用中，当x_0取在$\overline{x}$附近，n很大时，利用$\hat{y}_0 - y_0 \overset{近似}{\sim} N(0, \hat{\sigma}^{*2})$计算近似的边际误差$\delta^*$，此时$y_0$的置信水平为0.95的预测置信区间近似为$(\hat{y}-2\delta^*, \hat{y}+2\delta^*)$，置信水平为0.99的预测置信区间近似为$(\hat{y}-3\delta^*, \hat{y}+3\delta^*)$。

【例 7-2】 大家知道，营业税税收总额y与社会商品零售总额x有关。为了通过社会商品零售总额预测营业税税收总额，需要了解两者之间的关系。现收集了9组数据，见表7-3。

表7-3 社会商品零售总额与营业税税收总额 （单位：亿元）

序号	社会商品零售总额x	营业税税收总额y
1	142.08	3.93
2	177.30	5.96
3	204.68	7.85
4	242.88	9.82
5	316.24	12.50
6	341.99	15.55
7	332.69	15.79
8	389.29	16.39
9	453.40	18.45

试利用关于营业税税收额y与商品零售额x的回归方程，预测当前商品零售额x=300亿元时，营业税税收额y的值。

分析：进行点预测和区间预测。由于x=300亿元接近商品零售额的平均值，故用近似置信区间进行区间预测，显著性水平取α=0.05。

其实现的MATLAB程序代码如下：

```
>> clear all;
x=[142.08,177.30,204.68,242.88,316.24,341.99,332.69,389.29,453.40]';
y=[3.93,5.96,7.85,9.82,12.50,15.55,15.79,16.39,18.45]';
X=[ones(length(x),1),x];                    %构造自变量观测值矩阵
[b,bint,r,rint,stats]=regress(y,X);         %线性回归建模与评价
b,stats                                     %显示所关心的输出参数
x0=300;
y0=b(1)+b(2)*x0                             %点预测
SSE=sum((y-(b(1)+b(2)*x)).^2);              %计算残差平方和
STD=sqrt(SSE/(length(x)-2));                %计算标准误差
DELTA=2*STD;                                %计算0.05显著性水平下的边际误差
```

```
ci=[y0-DELTA,y0+DELTA]                    %0.95 置信区间
```

运行程序，输出如下：

```
b =
    -2.2610
     0.0487
stats =
     0.9625   179.7711     0.0000     1.1315
y0 =
    12.3423
ci =
    10.2149    14.4698
```

由此可知，回归方程为 $\hat{y}=-2.2610+0.0487x$，回归方程高度显著，可决系数 $r^2=0.9625$，模型方差的估计 $\hat{\sigma}^2=1.1315$。

即当社会商品零售总额为 300 亿元时，营业税平均税收总额的预测值约为 12.3423 亿元，其置信水平为 0.95 的置信区间为（10.2149 ,14.4698）。

7.1.5 一元非线性回归模型

在实际问题中，变量之间常常不是直线关系。这时，通常是选配一条比较接近的曲线，通过变量变换把非线性方程加以线性化，然后对线性化的方程应用最小二乘法求解回归方程。这就是本节要讲解的曲线回归问题。

最小二乘法的一个前提条件是函数 $y=f(x)$ 的具体类型已知，即要求首先确定 x 与 y 内在关系的函数类型。函数的类型可能是各种各样的，具体类型的确定或假设，一般有以下两个途径：一是根据有关的物理知识，确定两个变量之间的函数类型；二是把观测数据画在坐标纸上，将散点图与已知的函数曲线对比，选取最接近散点分布的曲线进行试算。

常见的一些非线性函数及线性化方法如下：

（1）倒幂函数 $y=a+b\dfrac{1}{x}$ 型

令 $x'=\dfrac{1}{x}$，则 $y=a+bx'$。

（2）双曲线 $\dfrac{1}{y}=a+b\dfrac{1}{x}$ 型

令 $y'=\dfrac{1}{y}$，$x'=\dfrac{1}{x}$，则 $y'=a+bx'$。

（3）幂函数曲线 $y=dx^b$ 型

令 $y'=\ln y$，$x'=\ln x$，$a=\ln d$，则 $y'=a+bx'$。

（4）指数曲线 $y=de^{bx}$ 型

令 $y'=\ln y$，$a=\ln d$，则 $y'=a+bx$。



（5）倒指数曲线 $y = de^{\frac{b}{x}}$ 型。

令 $y' = \ln y$，$x' = \frac{1}{x}$，$a = \ln d$，则 $y' = a + bx'$。

（6）对数曲线 $y = a + b\ln x$ 型

令 $x' = \ln x$，则 $y = a + bx'$

（7）S 型曲线 $y = \frac{1}{a + be^{-x}}$ 型

令 $y' = \frac{1}{y}$，$x' = e^{-x}$，则 $y' = a + bx'$。

综上所述，许多曲线都可以通过变换化为直线，于是可以按直线拟合的办法来处理。在线性化方法中，对数变换是常用的方法之一。当函数 $y = f(x)$ 的表达式不清楚时，往往可用对数变换进行试探看是否能线性化。通常把观测值标在对数坐标图中，当表现出良好的线性时，便可对变换后的数据进行回归分析，之后将得到的结果再代回原方程。因而，回归分析是对变换后的数据进行的，所得结果仅对变换后的数据来说是最佳拟合，当变换回原数据坐标时，所得的回归曲线，严格地说并不是最佳拟合，不过，其拟合程度通常是令人满意的。

进行对数变换时，必须使用原数据的实际观测值，而不能用经等差变换后的相对差值。例如，对原观测值 11 和 12 应用等差变换可以简化计算，用它们与 10 的相对差值（即 1 和 2）来描绘图形并不影响曲线的形状。然而，对数坐标中的距离代表的是比值，显然 11 和 12 的比同 1 和 2 的比是完全不同的。

可以看到，在所配曲线的回归中，可决系数 r^2、剩余标准误差 S_y、F 值的计算稍有不同。x'、y' 等仅仅是为了变量变换，使曲线方程变为直线方程，然而，要求的是所配曲线与观测数据拟合较好，所以计算 r^2、S_y、F 值时，应首先根据已建立的回归方程，用 x_i 依次代入，得到 y_i 后再计算残差平方和 $S_e = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ 及总平方和 $S_T = \sum_{i=1}^{n}(y_i - \overline{y})^2$，于是

$$r^2 = 1 - \frac{S_e}{S_T} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{7-29}$$

$$S_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}} \tag{7-30}$$

$$F = \frac{回归平方和/f_{回}}{残差平方和/f_{残}} = \frac{S_R/1}{S_e/n-2} \tag{7-31}$$

式中，$S_R = S_T - S_e$。

【例 7-3】 某雌性鱼的体长（cm）和体重（kg）的值见表 7-4，试对鱼的体重与体长进行回归分析。

表 7-4 雌性鱼的体长与体重的数据表

类别 \ 序号	1	2	3	4	5	6	7	8
体长 x/cm	70.70	98.25	112.57	122.48	138.46	148.00	152.00	162.00
体重 y/kg	1.00	4.85	6.59	9.01	12.34	15.50	21.25	22.11

分析：根据实际观测值在直角坐标系中作散点图，选定曲线类型，从散点图（见图 7-3）中实测点的分布趋势看出它比较接近幂函数曲线图形，因而选用 $y = ax^b$ 来进行拟合。由于是非线性回归，所以可用两种方法求出参数 a，b。一种是用 m 文件定义的非线性函数 $y = ax^b$，然后在主程序中使用非线性回归命令 nlinfit 求解。另一种是线性化，即将非线性模型转化成线性模型，只要对 $y = ax^b$ 取对数，即得 $\ln y = \ln a + b\ln x$，令 $y_1 = \ln y$，$a_1 = \ln a$，$x_1 = \ln x$，则得线性模型 $y_1 = a_1 + bx_1$。

其实现的 MATLAB 程序代码如下：

（第一种方法）首先定义非线性函数，并保存为 m 文件 yut.m。

```
function y=yut(beta,x)
y=beta(1)*x.^beta(2);
```

其实现的 MATLAB 程序代码如下：

```
>> %输入数据
x=[70.70,98.25,112.57,122.48,138.46,148.00,152.00,162.00];
y=[1.00,4.85,6.59,9.01,12.34,15.50,21.25,22.11];
beta0=[0.1,3]';
%求回归系数
[beta,r,J]=nlinfit(x',y','yut',beta0);
beta
%预测及作图
[YY,delta]=nlpredci('yut',x',beta,r,J);
plot(x,y,'k+',x,YY,'r');
```

运行程序，输出如下：

```
beta =
    0.000000758190151
    3.385125797710225
```

因为 MATLAB 默认是 short 型，其结果只保留 4 位小数，故这种情形下无法看出结果。把 MATLAB 设置为 long 型，输出结果如上。

结果表明，参数的估计值 $\hat{a} = 7.58\times10^{-7}$，$\hat{b} = 3.3851$，故回归模型为

$$y = 7.58\times10^{-7}x^{3.3851}$$

数据的散点图与回归线图如图 7-3 所示。从图 7-3 可看出，回归线能较好地表示散点图的形状，因此，回归模型成立。

（第二种方法）其实现的 MATLAB 程序代码如下：

```
>> %输入数据
x=[70.70,98.25,112.57,122.48,138.46,148.00,152.00,162.00];
y=[1.00,4.85,6.59,9.01,12.34,15.50,21.25,22.11];
```



```
%对数据作对数变换
x1=log(x);
y1=log(y);
%求线性回归系数
x2=[ones(8,1),x1'];
[b,bint,rint,stats]=regress(y1',x2);
b
a1=exp(b(1))
%预测及作图
z=b(1)+b(2)*x1;
yc=exp(z);  %体重的预测值
plot(x,y,'k+',x,yc,'r');
```

运行程序，输出如下：

```
b =
  -15.3913
    3.6494
a1 =
  2.0684e-007
```

结果表明，参数的估计值 $\hat{a}=2.068\times10^{-7}$，$\hat{b}=3.6491$，故回归模型为

$$y=2.068\times10^{-7}x^{3.6494}$$

数据的散点图与回归线图如图 7-3 所示。

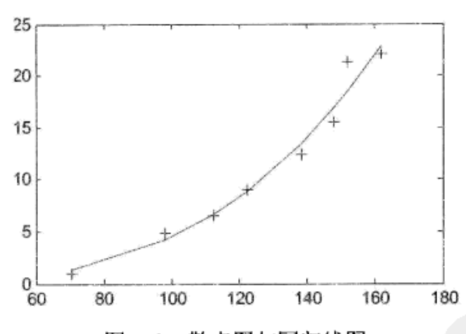


图 7-3 散点图与回归线图

比较两种方法，建立的回归模型有一定的差异，这是为什么呢？可以计算两种方法的残差平方和，第一种方法的残差平方和 s1=sum(r.^2)=12.1084，第二种方法的残差平方和 s2=sum((y-yc).^2)=14.0245，s2 大于 s1。一个合理的解释是：由于调用了不同的 MATLAB 命令，产生了计算误差，特别是，第二种方法对数据进行对数化变换可能造成更大的误差。

7.2 多元线性回归分析

一元线性回归将影响因变量的自变量限制为一个，这在现实中的大多社会经济现象中并不容易做到，因而应用回归分析时，常常要有更一般的模型，把两个或更多个解释变量的影

响分别估计在内，这就是多元回归，亦称为多重回归或复回归。当影响因素与因变量之间是线性关系时，所进行的回归分析就是多元线性回归。

7.2.1 多元线性回归分析的基本定义

在实际问题中，遇到更多的问题是讨论随机变量 y 与非随机变量 $x_1,x_2,\cdots,x_m$ 之间的关系。假设它们具有线性关系

$$y=\beta_0+\beta_1x_1+\cdots+\beta_mx_m+\varepsilon \tag{7-32}$$

式中，$\varepsilon\sim N(0,\sigma^2)$，$\beta_1,\beta_2,\cdots,\beta_m$，$\sigma^2$ 都是未知参数，一般称式（7-32）定义的模型为多元线性回归模型，$x_1,x_2,\cdots,x_m$ 为回归变量，$\beta_1,\beta_2,\cdots,\beta_m$ 为回归系数。

假设 $x_{1i},x_{2i},\cdots,x_{mi},y_i\ (i=1,2,\cdots,n)$ 是 $x_1,x_2,\cdots,x_m,y$ 的 n 个观测值，则它们满足关系

$$y_i=\beta_0+\beta_1x_{1i}+\cdots+\beta_{mi}x_{mi}+\varepsilon_i,\quad i=1,2,\cdots,n \tag{7-33}$$

式中，ε_i 相互独立，且 $\varepsilon_i\sim N(0,\sigma^2)$。

由于假设 ε_i 相互独立，则 y_i 也相互独立，且

$$\begin{cases}E\{y_i\}=\beta_0+\beta_1x_{1i}+\cdots+\beta_{mi}x_{mi}\\ D\{y_i\}=\sigma^2\end{cases} \tag{7-34}$$

7.2.2 矩阵表示法

要建立多元线性回归模型，首先要估计未知参数 $\beta_1,\beta_2,\cdots,\beta_m$，为此进行 $n(n\geqslant p)$ 次独立观测，得到 n 组数据（称为样本）

$$x_{i_1},x_{i_2},\cdots,x_{i_{m-1}},y_i,\quad i=1,2,\cdots,n$$

它们应满足式（7-33），即有

$$\begin{cases}y_1=\beta_0+\beta_1x_{11}+\beta_2x_{12}+\cdots+\beta_{m-1}x_{1m-1}+\varepsilon_1\\ y_2=\beta_0+\beta_1x_{21}+\beta_2x_{22}+\cdots+\beta_{m-1}x_{2m-1}+\varepsilon_2\\ \vdots\\ y_n=\beta_0+\beta_1x_{n1}+\beta_2x_{n2}+\cdots+\beta_{m-1}x_{nm-1}+\varepsilon_n\end{cases} \tag{7-35}$$

式中，$\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n$ 相互独立，且服从 $N(0,\sigma^2)$ 分布。令

$$\boldsymbol{Y}=\begin{pmatrix}y_1\\y_2\\\vdots\\y_n\end{pmatrix}_{n\times1},\quad \boldsymbol{X}=\begin{pmatrix}1&x_{11}&x_{12}&\cdots&x_{1m-1}\\1&x_{21}&x_{22}&\cdots&x_{2m-1}\\\vdots&\vdots&\vdots&&\vdots\\1&x_{n1}&x_{n2}&\cdots&x_{nm-1}\end{pmatrix}_{n\times m}$$

$$\boldsymbol{\beta}=\begin{pmatrix}\beta_0\\\beta_1\\\vdots\\\beta_{m-1}\end{pmatrix}_{m\times1},\quad \boldsymbol{\varepsilon}=\begin{pmatrix}\varepsilon_0\\\varepsilon_1\\\vdots\\\varepsilon_n\end{pmatrix}_{n\times1}$$

则式（7-35）可简写为如下形式

$$\begin{cases} \boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \\ \varepsilon \sim N(0, \sigma^2 I_n) \end{cases} \tag{7-36}$$



式中，$\boldsymbol{Y}$ 称为观测向量，$\boldsymbol{X}$ 称为设计矩阵，它们是由观测数据得到的，是已知的，并假定 $\boldsymbol{X}$ 为列满秩，即 $\text{rand}(\boldsymbol{X}) = m$；$\boldsymbol{\beta}$ 是待估计的未知参数向量；$\boldsymbol{\varepsilon}$ 是由不可观测的随机误差得到的。

式（7-36）称为多元线性回归模型的矩阵形式，亦称为高斯-马尔科夫线性模型，并简记为 $(\boldsymbol{Y}, \boldsymbol{X\beta}, \sigma^2 I_n)$。

对线性模型 $(\boldsymbol{Y}, \boldsymbol{X\beta}, \sigma^2 I_n)$ 所要考虑的问题主要是：

1）估计 β 与 σ^2，从而建立 y 与 $x_1, x_2, \cdots, x_{m-1}$ 的关系式。

2）对线性模型假设及 β 的某种假设进行检验。

3）对 y 进行预测及对自变量进行控制。

注意：假定 $n > m$。

7.2.3 未知参数估计

常常采用最小二乘法寻找 $\beta = (\beta_0, \beta_1, \beta_m)^{\mathrm{T}}$ 的估计值 $\hat{\beta}$，使得满足以下条件

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{n} x_{ji}\hat{\beta}_i)^2 = \min_{\beta} \sum_{i=1}^{n}(y_i - \sum_{j=0}^{m} x_{ji}\beta_i)^2 \tag{7-37}$$

利用微分法可以求解式（7-37），有

$$\sum_{i=1}^{n}(y_i - \sum_{j=0}^{m} x_{ji}\hat{\beta}_j)x_{ki} = 0, k = 0, 1, \cdots, m \tag{7-38}$$

式（7-38）变形为

$$\sum_{i=1}^{n} y_i x_{ki} = \sum_{i=1}^{n}\sum_{j=0}^{m} x_{ji}x_{ki}\hat{\beta}_j = \sum_{j=0}^{m}\left(\sum_{i=1}^{n} x_{ji}x_{ki}\right)\hat{\beta}_j \tag{7-39}$$

用矩阵表示为

$$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})\hat{\boldsymbol{\beta}} \tag{7-40}$$

可得

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y} \tag{7-41}$$

7.2.4 误差方差 σ^2 的估计

将自变量的各组观测值代入回归方程，可得因变量的各估计值（称为拟合值）为

$$\hat{\boldsymbol{Y}} \triangleq (\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_n) = \boldsymbol{X}\hat{\boldsymbol{\beta}}$$

称

$$\boldsymbol{e} \triangleq \boldsymbol{Y} - \hat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = \left[\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\right]\boldsymbol{Y} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} \tag{7-42}$$

为残差向量或剩余向量。

式中，$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}$ 为 n 阶幂等矩阵；$\boldsymbol{I}$ 为 n 阶单位矩阵。

称

$$Q_e=\boldsymbol{e}^{\mathrm{T}}\boldsymbol{e}=(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^{\mathrm{T}}(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})=\boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{Y}=\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}-\hat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$$

为剩余平方和。

由于 $E(\boldsymbol{Y})=\boldsymbol{X}\beta$ 且 $(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{Y}=0$，则

$$Q_e=\boldsymbol{e}^{\mathrm{T}}\boldsymbol{e}=(\boldsymbol{Y}-E(\boldsymbol{Y}))^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{H})(\boldsymbol{Y}-E(\boldsymbol{Y}))=\boldsymbol{\varepsilon}^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{\varepsilon}$$

由此可得

$$\begin{aligned}E(\boldsymbol{e}^{\mathrm{T}}\boldsymbol{e})&=E(\mathrm{tr}(\boldsymbol{\varepsilon}^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{H})\varepsilon))=\mathrm{tr}((\boldsymbol{I}-\boldsymbol{H})E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}))=\sigma^2\mathrm{tr}(\boldsymbol{I}-\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}})\\&=\sigma^2(n-\mathrm{tr}((\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}))=\sigma^2(n-m)\end{aligned}$$

其中，tr(*) 表示矩阵的迹。从而

$$\hat{\sigma}^2\triangleq\frac{1}{n-m}\mathrm{e}^{\mathrm{T}}\mathrm{e} \tag{7-43}$$

为 σ^2 的一个无偏估计。

7.2.5 有关的统计推断

1. 回归关系的统计推断

给定因变量 y 与自变量 $x_1,x_2,\cdots,x_{m-1}$ 的 n 组观测值，利用前述方法可得到未知参数 β 和 σ^2 的估计，从而可给出 y 与 $x_1,x_2,\cdots,x_{m-1}$ 之间的线性回归方程，但所求的回归方程是否有意义。也就是说，y 与 $x_1,x_2,\cdots,x_{m-1}$ 之间是否存在显著的线性关系，还需要对回归方程进行检验。

（1）建立方差分析表

- 离差平方和的分解。

观测值 $y_1,y_2,\cdots,y_n$ 之所以有差异，是由以下两个原因引起的。一方面是当 y 与 $x_1,x_2,\cdots,x_{m-1}$ 之间确有线性关系时，由于 $x_1,x_2,\cdots,x_{m-1}$ 取值的不同，而引起 y_i 值的变化；另一方面是除 y 与 $x_1,x_2,\cdots,x_{m-1}$ 的线性关系以外的因素，如 $x_1,x_2,\cdots,x_{m-1}$ 对 y 的非线性影响及随机因素的影响等。记 $\overline{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$，则数据的总的离差平方和

$$S_T\triangleq\frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^2 \tag{7-44}$$

反映了数据 $y_1,y_2,\cdots,y_n$ 波动性的大小。

残差平方和

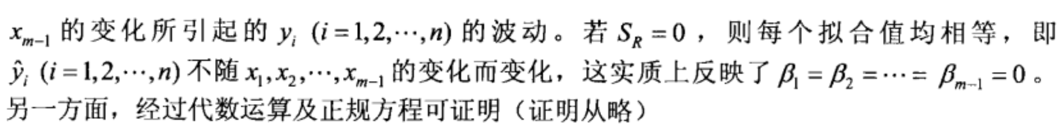
$$S_e=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \tag{7-45}$$

反映了除 y 与 $x_1,x_2,\cdots,x_{m-1}$ 的线性关系（即 $\hat{y}_i$）以外的因素引起的数据 $y_1,y_2,\cdots,y_n$ 的波动。若 $S_e=0$，则多个观测值可由线性关系精确拟合，S_e 越大，观测值和线性拟合之间的偏差也越大。

对于回归平方和

$$S_R = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 \tag{7-46}$$

可证明$\frac{1}{n}\sum_{i=1}^{n}\hat{y}_i = \overline{y}$，故$S_R$反映了线性拟合值与它们的平均值的总偏差，即由变量$x_1, x_2, \cdots,$



x_{m-1}的变化所引起的$y_i \ (i=1,2,\cdots,n)$的波动。若$S_R = 0$，则每个拟合值均相等，即$\hat{y}_i \ (i=1,2,\cdots,n)$不随$x_1, x_2, \cdots, x_{m-1}$的变化而变化，这实质上反映了$\beta_1 = \beta_2 = \cdots = \beta_{m-1} = 0$。另一方面，经过代数运算及正规方程可证明（证明从略）

$$S_T = S_R - S_e \tag{7-47}$$

因此，S_R越大，说明由线性回归关系所描述的$y_i \ (i=1,2,\cdots,n)$的波动性的比例就越大，即y与$x_1, x_2, \cdots, x_{m-1}$的线性关系就越显著。

另外，通过矩阵运算可证明S_T，S_e和S_R有如下形式的矩阵表示：

$$S_T = \sum_{i=1}^{n}(y_i - \overline{y})^2 = \boldsymbol{Y}^{\mathrm{T}}\left(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J}\right)\boldsymbol{Y} = \boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y} - \frac{1}{n}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{J}\boldsymbol{Y} \tag{7-48}$$

$$S_e = \mathrm{e}^{\mathrm{T}}\mathrm{e} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} = \boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y} - \hat{\beta}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y} \tag{7-49}$$

$$S_R = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 = \boldsymbol{Y}^{\mathrm{T}}\left(\mathrm{H} - \frac{1}{n}\boldsymbol{J}\right)\boldsymbol{Y} = \hat{\beta}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y} - \frac{1}{n}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{J}\boldsymbol{Y} \tag{7-50}$$

式中，$\boldsymbol{J}$表示一个元素全为1的n阶矩阵。$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}$为$n$阶对称幂等矩阵（可验证$\boldsymbol{H}^2 = \boldsymbol{H}$），$\boldsymbol{I}$为$\boldsymbol{n}$阶单位矩阵。

- 自由度的分解。

对于S_T（式（7-47）），其自由度也有相应的分解。这里的自由度是指平方和中独立变化项的数目。在S_T中，由于有一个关系式$S_T = \sum_{i=1}^{n}(y_i - \overline{y})^2 = 0$，即$y_i - \overline{y} \ (i=1,2,\cdots,n)$彼此不是独立变化的，故其自由度为$n-1$。

可以证明，S_e的自由度为$n-m$，S_R的自由度为$m-1$，因此对应于S_T的分解（式（7-47）），它们的自由度之间也有如下关系

$$n-1 = (n-m) + (m-1) \tag{7-51}$$

（2）线性回归关系的显著性检验

为检验y与$x_1, x_2, \cdots, x_{m-1}$之间是否存在显著的线性回归关系，即检验假设

$$\begin{cases} H_0：\beta_1 = \beta_2 = \cdots = \beta_{m-1} = 0 \\ H_1：至少有某一个\beta_i \neq 0, 1 \leqslant i \leqslant m-1 \end{cases} \tag{7-52}$$

这是因为若H_0成立，则$y = \beta_0 + \varepsilon$，即$y$与$x_1, x_2, \cdots, x_{m-1}$之间不存在线性回归关系。基于上述分析，构造如下检验统计量

$$F \triangleq \frac{V_R}{V_e} \tag{7-53}$$

当 H_0 为真时，可以证明 $F \sim F(m-1,n-m)$。由上述对回归平方和 S_R 的讲解可知，若 H_0 为假，则 F 的值有偏大的趋势。因此，给定显著性水平 α，查 F 分布表得临界值 $F_\alpha(m-1,n-m)$，计算 F 的观测值 F_0，若 $F_0 < F_\alpha(m-1,n-m)$，接受 H_0，即认为 y 与 $x_1,x_2,\cdots,x_{m-1}$ 之间存在显著的线性回归关系。

（3）拟合优度的测定——相关系数法

和一元线性回归分析类似，多元回归也可以用一个"相关系数" R 来衡量，即用回归平方和 S_R 在总平方和 S_T 中的比例来衡量，用 R 代替 r

$$R = \sqrt{\frac{S_R}{S_T}} \tag{7-54}$$

称为相关系数。它的意义和一元的相关系数 r 一样，$0 \leqslant R \leqslant 1$。

回归方程的精度用剩余标准差来表示

$$S = \sqrt{\frac{S_e}{n-m}} \tag{7-55}$$

注意：当作了整个回归方程分析的 F 检验后，就不必再作多相关系数的显著性研究了，它们实质上是等价的。

2. 回归参数的统计推断（偏回归系数检验）

回归关系显著并不意味着每个自变量 x_i $(1 \leqslant i \leqslant m-1)$ 对 y 的影响都显著，可能其中的某个或某些自变量对 y 的影响不显著。一般来说，总希望从回归方程中剔除对 y 的影响不显著的自变量，从而建立一个较为简单有效的回归方程，以便于实际应用。因为当一个回归方程包含有不显著的变量时，它不仅对利用回归方程作预测和控制带来麻烦，而且还会增大 $\hat{y}$ 的方差，从而影响预测的精度。为此，需要对每一个回归系数作显著性检验，显然，若某个自变量 x_i 对 y 无影响，那么在线性模型中，它的系数 β_j 应为零。因此，检验 x_i 的影响是否显著等价于检验假设

$$H_0:\ \beta_j = 0;\quad H_1:\ \beta_j \neq 0 \tag{7-56}$$

下面讲解此假设的检验问题。

设 $\hat{\beta} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$ 为 β 的最小二乘估计，则 $E(\hat{\beta}) = \beta$，因此，$\hat{\beta}$ 的协方差矩阵为

$$\mathrm{Cov}(\hat{\beta},\hat{\beta}) = D(\hat{\beta}) = E(\hat{\beta}-\beta)(\hat{\beta}-\beta)^{\mathrm{T}}$$

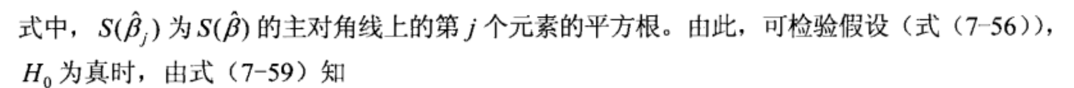
因为

$$E(\hat{\beta}) = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}E(Y) = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\beta = \beta$$

所以

$$\begin{aligned}\mathrm{Cov}(\hat{\beta},\hat{\beta}) &= E\{[(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y} - \boldsymbol{E}(\hat{\beta})][(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y} - E(\hat{\beta})]^{\mathrm{T}}\} \\ &= (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}E\{[\boldsymbol{Y}-\boldsymbol{X}\beta][\boldsymbol{Y}-\boldsymbol{X}\beta]^{\mathrm{T}}\} - [(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}]^{\mathrm{T}} \\ &= (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\sigma^2 I_n\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1} \\ &= \sigma^2(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\end{aligned} \tag{7-57}$$

由式（7-43）、式（7-49）知，$\dfrac{S_e}{n-m} \triangleq V_e$ 为 $\hat{\sigma}^2$ 的无偏估计，即 $\hat{\sigma}^2 = V_e$，因此以





$$S(\hat{\beta}) \triangleq V_e(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1} \tag{7-58}$$

作为$D(\hat{\beta})$的估计。可以证明

$$\frac{\hat{\beta}_j - \beta_j}{S(\hat{\beta}_j)} \sim t(n-m), \quad j = 1,2,\cdots,m-1 \tag{7-59}$$

式中，$S(\hat{\beta}_j)$为$S(\hat{\beta})$的主对角线上的第j个元素的平方根。由此，可检验假设（式（7-56）），H_0为真时，由式（7-59）知

$$t = \frac{\hat{\beta}_j}{S(\hat{\beta}_j)} \sim t(n-m) \tag{7-60}$$

若H_0为假，由于$E(\hat{\beta}_j) = \beta_j \neq 0$，则$|t|$有偏大的趋势。在显著性水平$\alpha$下，查表得$t_{\frac{\alpha}{2}}(n-m)$，记$t$的观测值为$t_0$，检验准则为

$$\begin{cases} 若 |t_0| < t_{\frac{\alpha}{2}}(n-m), & 则接受 H_0 \\ 若 |t_0| \geqslant t_{\frac{\alpha}{2}}(n-m), & 则拒绝 H_0 \end{cases}$$

另外，由式（7-60）可求得β_j的置信水平为$1-\alpha$的置信区间为

$$\hat{\beta}_j \pm t_{\frac{\alpha}{2}}(n-m)S(\hat{\beta}_j) \tag{7-61}$$

3. 关于预报值的统计推断

建立回归方程除了了解y与$x_1,x_2,\cdots,x_{m-1}$的相依关系外，另一个重要应用就是进行预报。

设给定了自变量的一组新观测值$x_{01},x_{02},\cdots,x_{0m-1}$，利用回归方程可设因变量$y$的预报值

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \hat{\beta}_2 x_{02} + \cdots + \hat{\beta}_{m-1} x_{0m-1} \tag{7-62}$$

$\hat{y}_0$实际上是对应于$x_{01},x_{02},\cdots,x_{0m-1}$的$y$的一个点估计。在实际应用中，更感兴趣的是给出$y$的真值$y_0$的区间估计。可以证明

$$\frac{\hat{y}_0 - y_0}{S(\hat{y}_0)} \sim t(n-m) \tag{7-63}$$

式中，

$$S(\hat{y}_0) = V_e[1 + \boldsymbol{X}_{new}^{\mathrm{T}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}_{new}]。 \tag{7-64}$$

式中，$\boldsymbol{X}_{new} = (1, x_{01}, x_{02}, \cdots, x_{0m-1})^{\mathrm{T}}$。由此，可设$y_0$的一个置信水平为$1-\alpha$的置信区间为

$$\hat{y}_0 \pm t_{\frac{\alpha}{2}}(n-m)S(\hat{y}_0) \tag{7-65}$$

【例 7-4】 某厂生产一种商品的销售量y与竞争对手的价格x_1、本厂的价格x_2有关，其销售记录见表 7-5。试根据这些数据建立y与x_1、x_2的关系式，对得到的模型和系数进行检验。

表 7-5 销售量与价格统计表

序　号	1	2	3	4	5	6	7	8	9	10
x_1/(元/件)	120	140	190	130	155	175	125	145	180	150
x_2/(元/件)	100	110	90	150	210	150	250	270	300	250
y/件	102	100	120	77	46	93	26	69	65	85

分析：为了确定一种商品的销售量与价格之间的关系，分别作出 y 与 x_1、x_2 的散点图（为二元线性散点图）。散点图显示它们之间近似为线性关系，因此可设定 y 与 x_1、x_2 的关系为二元线性回归模型：$y=\beta_0+\beta_1x_1+\beta_2x_2$。

其实现的 MATLAB 程序代码如下：

```
>> clear all;
%输入数据并作散点图(见图 7-4)
x1=[120   140   190   130   155   175   125   145   180   150]';
x2=[100   110   90    150   210   150   250   270   300   250]';
y=[102    100   120   77    46    93    26    69    65    85]';
figure;
plot(x1,y,'or',x2,y,'+');
%作二元线性回归
x=[ones(10,1),x1,x2];
[b,bint,r,rint,stats]=regress(y,x);
b,bint,stats,
%作残差分析图(见图 7-5)
figure; rcoplot(r,rint);
```

运行程序，输出如下：

```
b =
    66.5176
     0.4139
    -0.2698
bint =
   -32.5060   165.5411
    -0.2018     1.0296
    -0.4611    -0.0785
stats =
     0.6527     6.5786     0.0247   351.0445
```

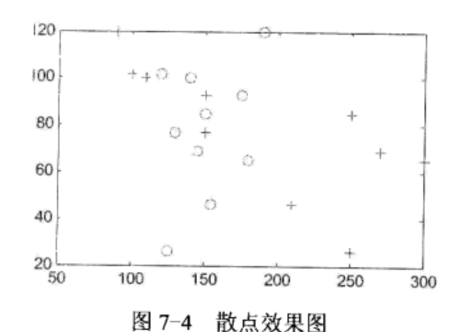


图 7-4 散点效果图

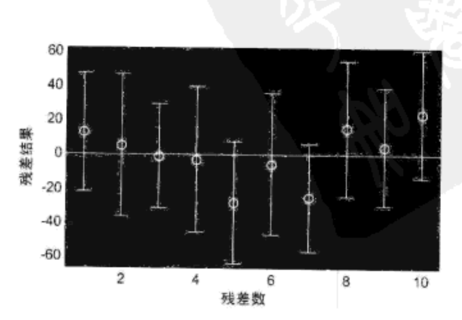


图 7-5 残差图

结果表明，线性回归方程为 $\hat{y}=66.5176+0.4139x_1-0.2698x_2$，可决系数 $r^2=0.6527$，$p=0.0247<0.05$，故回归模型成立。

【例 7-5】 某销售公司将其连续 18 个月的库存占用资金情况、广告投入的费用、员工薪酬及销售额等方面的数据作了汇总（见表 7-6）。该公司的管理人员试图根据这些数据找到销售额与其他 3 个变量之间的关系，以便进行销售额预测并为未来的工作决策提供参考依据。

1）试建立销售额的回归模型。

2）如果未来某月的库存占用资金为 150 万元，广告投入预算为 45 万元，员工薪酬总额为 27 万元，试根据建立的回归模型预测该月的销售额。

表 7-6 库存占用资金、广告投入、员工薪酬、销售额 （单位：万元）

月　份	库存占用资金 x_1	广告投入 x_2	员工薪酬 x_3	销售额 y
1	75.2	30.6	21.1	1090.4
2	77.6	31.3	21.4	1133
3	80.7	33.9	22.9	1242.1
4	76	29.6	21.4	1003.2
5	79.5	32.5	21.5	1283.2
6	81.8	27.9	21.7	1012.2
7	98.3	24.8	21.5	1098.8
8	67.7	23.6	21	826.3
9	74	33.9	22.4	1003.3
10	151	27.7	24.7	1554.6
11	90.8	45.5	23.2	1199
12	102.3	42.6	24.3	1483.1
13	115.6	40	23.1	1407.1
14	125	45.8	29.1	1551.3
15	137.8	51.7	24.6	1601.2
16	175.6	67.2	27.5	2311.7
17	155.2	65	26.5	2126.7
18	174.3	65.4	26.8	2256.5

分析：为了确定销售额 y 与库存占用资金 x_1、广告投入 x_2、员工薪酬 x_3 之间的关系，分别作出 y 与 x_1，y 与 x_2，y 与 x_3 的散点图。散点图显示它们之间近似为线性关系，因此可设定 y 与 x_1，x_2，x_3 的关系为三元线性回归模型：$y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3$。

其实现的 MATLAB 程序代码如下：

```
>> clear all;
%输入数据并作散点图
A=[75.2 30.6 21.1 1090.4;77.6 31.3 21.4    1133;80.7 33.9 22.9 1242.1;76 29.6 21.4      1003.2;...
    79.5 32.5 21.5 1283.2;81.8 27.9 21.7 1012.2;98.3        24.8 21.5 1098.8;67.7 23.6 21 826.3;...
    74 33.9 22.4     1003.3;151 27.7          24.7    1554.6;90.8    45.5    23.2    1199;102.3    42.6    24.3
```

```
    1483.1;...
    115.6 40    23.1 1407.1;125    45.8 29.1 1551.3;137.8 51.7    24.6 1601.2;175.6 67.2 27.5
    2311.7;...
    155.2 65    26.5 2126.7;174.3 65.4 26.8    2256.5];
figure;subplot(221);
plot(A(:,1),A(:,4),'*');title('销售额与库存占用资金');
subplot(222);
plot(A(:,2),A(:,4),'o');title('销售额与广告投入');
subplot(212);
plot(A(:,3),A(:,4),'+');title('销售额与员工薪酬总额');
%作多元回归
x=[ones(18,1) A(:,1:3)];
[b,bint,r,rint,stats]=regress(A(:,4),x);
b,bint,stats,
%预测
x1=[1 150 45 27];
y1=x1*b
%作残差分析图
figure(2);
rcoplot(r,rint);
```

运行程序，输出如下：

```
b =
  162.0632
    7.2739
   13.9575
   -4.3996
bint =
 -580.3603   904.4867
    4.3734    10.1743
    7.1649    20.7501
  -46.7796    37.9805
stats =
  1.0e+004 *
    0.0001    0.0105    0.0000    1.0078
y1 =
  1.7624e+003
```

结果表明，系数 $\beta_0 = 162.0632$，$\beta_1 = 7.2739$，$\beta_2 = 13.9575$，$\beta_3 = -4.3996$，且 β_0，β_1，β_2，β_3 在置信水平为 0.95 下的置信区间分别为[−580.3603 , 904.4867]、[4.3734 ,10.1743]、[7.1649, 20.7501]、[−46.7796 ,37.9805]，可决系数 $r^2 = 0.0001$，$p = 0.0000<0.05$，故回归模型

$$\hat{y} = 162.0632 + 7.2739x_1 + 13.9575x_2 - 4.3996x_3$$

成立。当未来某月的库存占用资金为 150 万元，广告投入预算为 45 万元，员工薪酬总额为 27 万元时，由模型预测该月的销售额为 1762.4 万元。

数据的散点图及回归模型的残差分析图如图 7-6、图 7-7 所示。

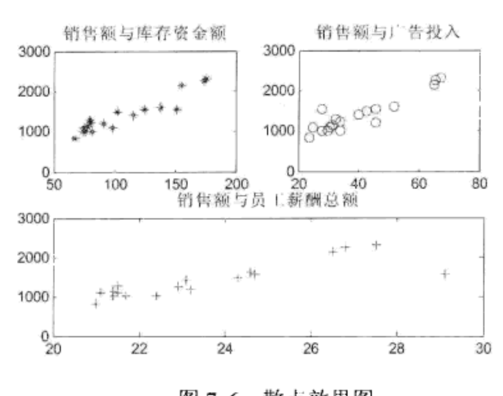



图 7-6 散点效果图

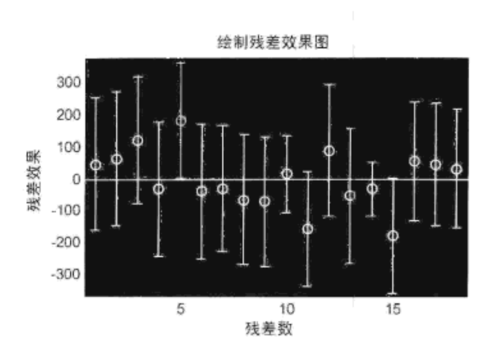


图 7-7 残差分析图



【例 7-6】 表 7-7 是血压与年龄、体重指数、吸烟习惯的统计数据。其中，吸烟习惯用 0 表示不吸烟，1 表示吸烟；体重指数=（体重（kg）/身高（m））的平方。试建立回归分析模型，分析血压与年龄、体重指数、吸烟习惯的关系。

表 7-7 血压与年龄、体重指数、吸烟习惯的统计数据

序号	血压/mmHg	年龄/岁	体重指数	吸烟习惯	序号	血压/mmHg	年龄/岁	体重指数	吸烟习惯
1	144	39	24.2	0	16	130	48	22.2	1
2	215	47	31.1	1	17	135	45	27.4	0
3	138	45	22.6	0	18	114	18	18.8	0
4	145	47	24.0	1	19	116	20	22.6	0
5	162	65	25.9	1	20	124	19	21.5	0
6	142	46	25.1	0	21	136	36	25.0	0
7	170	67	29.5	1	22	142	50	26.2	1
8	124	42	19.7	0	23	120	39	23.5	0
9	158	67	27.2	1	24	120	21	20.3	0
10	154	56	19.3	0	25	160	44	27.1	1
11	162	64	28.0	1	26	158	53	28.6	1
12	150	56	25.8	0	27	144	63	28.3	0
13	140	59	27.3	0	28	130	29	22	1
14	110	34	20.1	0	29	125	25	25.3	0
15	128	42	21.7	0	30	175	69	27.4	0

分析：为了确定血压与上述 3 个指标之间存在何种关系，首先作出血压与年龄，血压与体重指数之间的散点图，如图 7-8 和图 7-9 所示。

其实现的 MATLAB 程序代码如下：

```
A=[144,39,24.2,0;215,47,31.1,1;138,45,22.6,0;145,47,24.0,1;162,65,25.9,1;142,46,25.1,0;...
   170,67,29.5,1;124,42,19.7,0;158,67,27.2,1;154,56,19.3,0;162,64,28.0,1;150,56,25.8,0;...
```

```
    140,59,27.3,0;110,34,20.1,0;128,42,21.7,0;130,48,22.2,1;135,45,27.4,0;114,18,18.8,0;...
    116,20,22.6,0;124,19,21.5,0;136,36,25.0,0;142,50,26.2,1;120,39,23.5,0;120,21,20.3,0;...
    160,44,27.1,1;158,53,28.6,1;144,63,28.3,0;130,29,22,1;125,25,25.3,0;175,69,27.4,0];
figure;
plot(A(:,1),A(:,2),'*');title('血压与年龄的散点图');
figure(2);
plot(A(:,1),A(:,3),'o');title('血压与体重指数的散点图');
%作多元回归
x=[ones(30,1) A(:,2:4)];
[b,bint,r,rint,stats]=regress(A(:,1),x);
b,bint,stats,
```

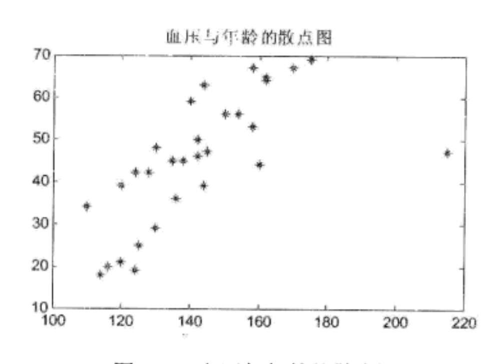


图 7-8　血压与年龄的散点图

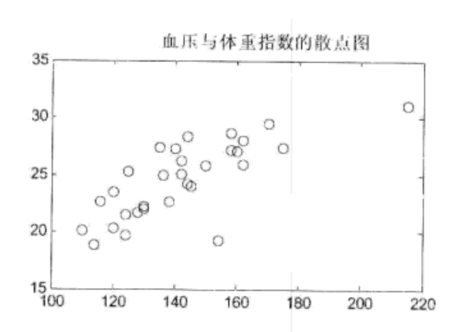


图 7-9　血压与体重指数的散点图

从图中可以看出以下几点：

1）随着年龄的增长血压有增高的趋势；随着体重指数的增长，血压也有增高的趋势。

2）从总体上看，血压与年龄、血压与体重指数存在一定的线性相关性，所以可建立多元线性回归模型：

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

式中，回归系数 $\beta_0, \beta_1, \beta_2, \beta_3$ 由数据估计，ε 是随机误差。

其次，求出回归系数 $\beta_0, \beta_1, \beta_2, \beta_3$ 的估计值与置信区间，并求出相应的统计量，所得结果见表 7-8。

表 7-8　回归模型的系数、系数置信区间与统计量

回 归 系 数	回归系数估计值	回归系数置信区间
β_0	41.0079	[−1.5461　83.5619]
β_1	0.4220	[−0.0173　0.8612]
β_2	3.2142	[1.1132　5.3153]
β_3	8.8929	[−2.9765　20.7623]
$r^2 = 0.6661$，$F = 17.2900$，$p < 0.0001$，$\sigma^2 = 180.2614$		

从表 7-8 可知，由于 β_1，β_3 的置信区间包含零点，因此模型需要改进，为此作出残差与残差置信区间的图形（见图 7-10）。

此时，从图形可见到第二个点与第十个点是异常的，剔除这两点，再次进行回归，得到改进数据后的回归模型的系数、系数置信区间与统计量结果，见表 7-9。

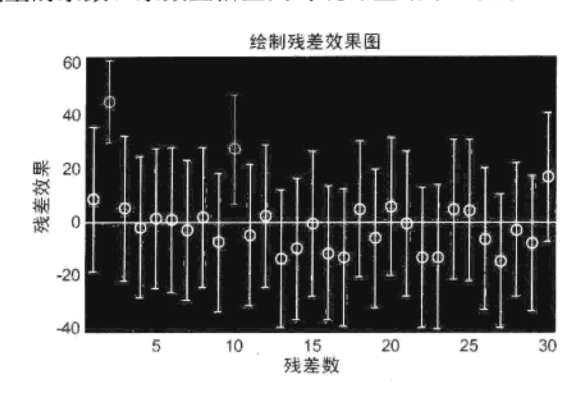




图 7-10 残差图

表 7-9 改进后的回归模型的系数、系数置信区间与统计量

回归系数	回归系数估计值	回归系数置信区间
β_0	58.5101	[29.9064 87.1138]
β_1	0.4303	[0.1273 0.7332]
β_2	2.3449	[0.8509 3.8389]
β_3	10.3065	[3.3878 17.2253]
$r^2 = 0.8462$, $F = 44.0087$, $p < 0.0001$, $\sigma^2 = 53.6604$		

从表中可知，这时的所有参数置信区间不包含零点，F 统计量增大，可决系数从 0.6855 增大到 0.8462，得到回归模型为

$$\hat{y} = 58.5101 + 0.4303x_1 + 2.3449x_2 + 10.3065x_3$$

最后，对模型进行检验，说明模型的合理性。

1）残差的正态检验。由 jbtest 检验，h=0 表明残差服从正态分布，进而由 t 检验可知 $h_1 = 0$，$p_1 = 1$，故残差服从均值为零的正态分布。

2）残差的异方差检验，也称为戈德菲尔德-匡特（Goldfeld-Quant）检验。

将 28 个数据按从小到大的顺序排列，去掉中间的 6 个数据，得到 F 统计量的观测值为 $f = 1.6604$，由 $F(7,7) = 3.79$，可知 $f = 1.6604 < 3.79$，故不存在异方差。

3）残差的自相关性检验，也称为 D-W 检验。

通过计算 $DW = \dfrac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$，得到 $DW = 1.4330$，查表后 dl=0.97，du=1.41，由于 1.41=du<DW=1.4330<4−du=2.59，可知残差不存在自相关性。

其实现的 MATLAB 程序代码如下：

```
A=[144,39,24.2,0;215,47,31.1,1;138,45,22.6,0;145,47,24.0,1;162,65,25.9,1;142,46,25.1,0;...
    170,67,29.5,1;124,42,19.7,0;158,67,27.2,1;154,56,19.3,0;162,64,28.0,1;150,56,25.8,0;...
    140,59,27.3,0;110,34,20.1,0;128,42,21.7,0;130,48,22.2,1;135,45,27.4,0;114,18,18.8,0;...
    116,20,22.6,0;124,19,21.5,0;136,36,25.0,0;142,50,26.2,1;120,39,23.5,0;120,21,20.3,0;...
    160,44,27.1,1;158,53,28.6,1;144,63,28.3,0;130,29,22,1;125,25,25.3,0;175,69,27.4,0];
%求多元回归的参数估计
[b,bint,r,rint,s]=regress(A(:,1),[ones(30,1),A(:,2:4)]);
s2=sum(r.^2)/(30-3-1)    %计算残差平方和
b,bint,s      %显示结果
rcoplot(r,rint)    %作残差与残差置信区间的图形
%剔除异常点并执行回归程序
A1=A([1,3:9,11:30],:);    %剔除异常点
[b2,bint2,r2,rint2,s1]=regress(A1(:,1),[ones(28,1),A1(:,2:4)])
%残差检验程序
[h,p]=jbtest(r2)    %正态性检验
[h1,p1]=ttest(r2,0)   %t 检验
%异方差检验
[c,i]=sort(A1(:,1));   %将样本值按被解释变量从小到大的顺序排序
A2=A1(i,2:4);
%在所有样本点中删去中间的 6 个点,将余下的点分为两组,取前 11 个点作回归
[b10,bint10,r10,rint10,s10]=regress(c(1:11),[ones(11,1),A2(1:11,:)]);
[b1h,bint1h,r1h,rint1h,s1h]=regress(c(18:28),[ones(11,1),A2(18:28,:)]); %取后 11 个点作回归
yf1=sum(r1h.^2)/sum(r10.^2)   %计算 F 检验统计量值
%自相关性检验
dw=sum(diff(r2).^2)/sum(r2.^2)    %计算 DW 统计量
```

1）回归模型的基本假定：利用样本数据估计回归模型中的参数，为了选择适当的参数估计方法，提高估计的精度，通常需要事先对模型的随机误差项和解释变量的特性进行假设。

假设 1 解释变量是非随机的或固定的，且各 X 之间互不相关（无多重共线性）。

假设 2 随机误差项具有零均值、同方差及序列不相关性，即

$$E(\varepsilon_i)=0\text{，}V=E(\varepsilon_i^2)=\sigma^2\text{，}\operatorname{Cov}(\varepsilon_i,\varepsilon_j)=0\text{，}i\neq j;i,j=1,2,\cdots,n$$

假设 3 解释变量与随机项不相关 $\operatorname{Cov}(\varepsilon_i,\varepsilon_j)=0$。

假设 4 随机误差项满足正态分布 $\varepsilon_i \sim N(0,\sigma^2)$。

将满足这些假设的回归模型称为古典回归模型。直观地看，这些假设的作用是便于分离回归模型中每个因素的单独影响，在回归分析的参数估计和统计检验理论中，许多结论都以这些假设作为基础。换句话说，这些假设的成立与否将直接影响回归分析中统计推断的结论。

2）对于实际问题，在建立模型时应注意以下问题：

- 模型中是否应该具有常数项，这取决于该常数的实际意义。
- 对于涉及有关专业的问题，需请教有关专家决定自变量的取舍。对于本题的结果，医学专家认为模型中的常数无法给出合理的解释，此外吸烟与血压的高低没有关系。

因此可以考虑建立血压与年龄、体重指数之间的二元回归模型。

【例 7-7】 假设线性回归方程为 $y = x_1 - 1.232x_2 + 2.23x_3 + 2x_4 + 4x_5 + 3.792x_6$，试生成 120 组随机输入值 x_i，计算出输出向量 $\boldsymbol{Y}$。以这些信息为已知，观察是否能由最小二乘法得出待定系数 a_i 的估计值，并得出置信区间。



其实现的 MATLAB 程序代码如下：

```
>> a=[1,-1.232,2.23,2,4,3.792]';
>> X=randn(120,6);
>> y=X*a;  .
>> a1=inv(X'*X)*X'*y
```

运行程序，输出如下：

```
a1 =
      1.0000
     -1.2320
      2.2300
      2.0000
      4.0000
      3.7920
```

可见，因为输出值完全由精确计算得出，所以线性回归参数估计的误差是极其微小的，可以忽略。用 regress 函数还可以计算出置信水平为 0.98 的置信区间。

```
>> [a,aint]=regress(y,X,0.02)
```

运行程序，输出如下：

```
a =
      1.0000
     -1.2320
      2.2300
      2.0000
      4.0000
      3.7920
aint =
      1.0000      1.0000
     -1.2320     -1.2320
      2.2300      2.2300
      2.0000      2.0000
      4.0000      4.0000
      3.7920      3.7920
```

假设观测的输出数据样本噪声污染，则可以给出输出数据样本叠加上 $N(0,0.5)$ 区间的正态分布噪声，这时可以用下面的语句进行线性回归分析，得出待定系数向量的估计参数及置信区间，通过 errorbar 函数还可以计算出图形绘制参数估计的置信区间，如图 7-11a 所示。

其实现的 MATLAB 程序代码如下：

```
>> yhat=y+sqrt(0.5)*randn(120,1);
>> [a,aint]=regress(yhat,X,0.02)
```

运行程序，输出如下：

```
a = 1.0974
   -1.1740
    2.2491
    2.0168
    4.0086
    3.8107
aint =
    0.9304    1.2645
   -1.3299   -1.0181
    2.0914    2.4067
    1.8694    2.1643
    3.8553    4.1619
    3.6488    3.9727

>> errorbar(1:6,a,aint(:,1)-a,aint(:,2)-a)
```

所以减小噪声的方差，假设方差为 0.1，则可以得出新噪声下参数估计的结果，如图 7-11b 所示。显然，估计出的参数更精确。

其实现的 MATLAB 程序代码如下：

```
>> yhat=y+sqrt(0.1)*randn(120,1);
>> [a,aint]=regress(yhat,X,0.02);
>> errorbar(1:6,a,aint(:,1)-a,aint(:,2)-a)
```

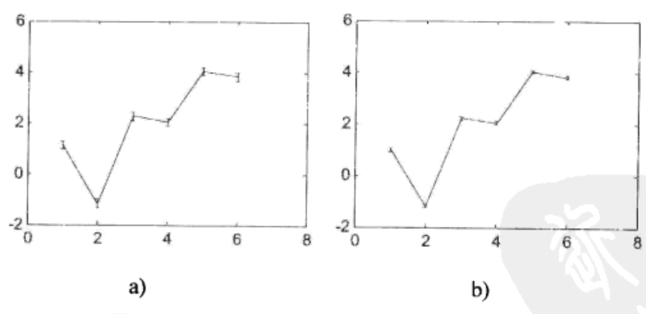


图 7-11 参数估计及置信区间图形表示

a) 噪声 $\sigma^2=0.5$　b) 噪声 $\sigma^2=0.1$

7.3 偏最小二乘回归分析

经典多元线性回归分析（MLR）是研究变量之间的相关关系的基本方法。但是，下面两个问题制约着其应用的效能：一是样本容量要求很高，一般应大于 30 或大于自变量数的

5~10 倍；二是消除变量间多重相关性很难。若在变量之间存在严重的多重相关性，将对回归建模与模型分析工作带来如下危害。

- 在自变量之间存在严重的多重相关性的情况下，将造成回归资料矩阵的严重病态性，进而使模型参数的最小二乘估计失真。回归系数的估计方差将随着自变量之间相关程度的不断增强而迅速扩大，回归系数的估计值对样本数据的微小变化变得非常敏感，回归系数估计值的稳定性将变得很差。
- 在自变量高度相关的条件下，用最小二乘法得到的回归模型其回归系数的物理含义很难解释。许多从专业知识上看似乎十分重要的变量，其回归系数的取值变得微不足道，甚至还会出现回归系数的符号与人们的实际概念完全相反的现象。
- 存在严重的多重共线性影响时，回归系数的统计检验将难以通过。



回归建模过程中必须要解决多重共线性问题。常见的方法是用逐步回归法来进行变量的筛选，去掉不太重要的相关性变量。然而，逐步回归法存在下列问题：一是缺乏对变量间多重相关性进行判定的十分可靠的检验方法；二是删除部分多重相关变量的做法常导致增大模型的解释误差，将本应保留的系数信息舍弃，使得接受错误结论的可能及做出错误决策的风险不断增长。

在克服变量多重相关性对系统回归建模干扰的过程中，1983 年，瑞典的 S.Wold 和 C.Albano 等人提出了偏最小二乘回归分析（PLS）方法。PLS 方法开辟了一种有效的技术途径，在处理样本容量小、解释变量个数多、变量间存在严重的多重相关性问题方面具有独特的优势，并且可以同时实现回归模型、数据结构简化及两组变量间的相关分析。

7.3.1 偏最小二乘回归方法的数据结构与建模思想

设有 q 个因变量 $y_1, y_2, \cdots, y_q$ 与 p 个自变量 $x_1, x_2, \cdots, x_p$，为了研究因变量与自变量的统计关系，观测了 n 个样本点，由此分别构成了自变量与因变量的“样本点×变量”型的数据矩阵，记为

$$\boldsymbol{X} = (x_{ij})_{n \times p} = (x_1, x_2, \cdots, x_p)$$

和

$$\boldsymbol{Y} = (Y_{ij})_{n \times q} = (y_1, y_2, \cdots, y_q)$$

PLS 方法在建模过程中采用了信息综合与筛选技术，不直接考虑因变量系统 Y 对自变量系统 X 的回归模型，而是从自变量系统 X 中逐步提取 m 个对自变量系统 X 和因变量系统 Y 都具有最佳解释能力的新综合变量 $t_1, t_2, \cdots, t_m$ $(m \leqslant p)$，亦称为主成分。首先建立 y_k 对主成分 $t_1, t_2, \cdots, t_m$ 的 MLR 回归方程，然后还原为 y_k 关于原自变量系统 $x_1, x_2, \cdots, x_p$ 的 PLS 回归方程，其中 $k = 1, 2, \cdots, q$。

PLS 方法的关键性技术是提取主成分，基本思想如下。

第一步，分别在 X 和 Y 中提取第一主成分 t_1 和 u_1，并且要求：

① 主成分的代表性，t_1 和 u_1 应尽可能多地携带各自变量系统中的变异信息。

② 主成分的相关性，t_1 和 u_1 的相关程度能够达到最大，即 t_1 对因变量系统有很强的解释能力。

这两个要求表明，PLS 方法主成分的提取与主成分分析中主成分的提取既有相似之处（代表性要求），又有不同的地方（相关性要求）。

第二步，在第一个主成分 t_1 和 u_1 被提取后，分别实施

① 各自变量对自变量系统第一主成分的回归（即用 t_1 表示 X ）。

② 各因变量对自变量系统第一主分成的回归（即用 t_1 表示 Y ）。

如果回归方程已经达到满意的精度，则算法终止；否则，将利用 X 被 t_1 解释后的残余信息，以及 Y 被 t_1 解释后的残余信息进行第二轮的成分提取。如此往复，直到达到一个较满意的精度为止。

7.3.2 偏最小二乘回归方法的算法步骤

首先要进行预备分析，目的是判断自变量（因变量）是否存在多重相关性，判断因变量与自变量是否存在相关关系，进而决定是否需要采用 PLS 方法建模。具体计算方法是：记矩阵 $\boldsymbol{Z}=(\boldsymbol{X},\boldsymbol{Y})$，求 $\boldsymbol{Z}$ 的各列数据之间的简单相关系数；然后，按下列步骤建立偏最小二乘回归方程。

1. 标准化原始数据

标准化后的数据矩阵记为 $\boldsymbol{E}_0=(e_{ij})_{n\times p}$ 和 $\boldsymbol{F}_0=(f_{ij})_{n\times q}$，其中

$$e_{ij}=\frac{x_{ij}-\overline{x}_j}{sx_j},\quad i=1,2,\cdots,n;j=1,2,\cdots,p \tag{7-66}$$

$$f_{ij}=\frac{y_{ij}-\overline{y}_j}{sy_j},\quad i=1,2,\cdots,n;j=1,2,\cdots,q \tag{7-67}$$

式（7-66）和式（7-67）中，$\overline{x}_j$，$\overline{y}_j$ 分别为矩阵 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 的第 j 列数据的平均值；sx_j，sy_j 分别为矩阵 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 的第 j 列数据的标准差。

2. 建立回归方程

（1）建立关于主成分的 MLR 回归方程

求出 $\boldsymbol{F}_0$ 在 $t_1,t_2,\cdots,t_m$ 上的 MLR 回归方程

$$\boldsymbol{F}_0=t_1\boldsymbol{r}_1^{\mathrm{T}}+t_2\boldsymbol{r}_2^{\mathrm{T}}+\cdots+t_m\boldsymbol{r}_m^{\mathrm{T}}+\boldsymbol{F}_m \tag{7-68}$$

（2）变换为关于标准化变量的 PLS 回归方程

将 $t_i=\boldsymbol{E}_{i-1}w_i=\boldsymbol{E}_0w_i^*(i=1,2,\cdots,m)$ 代入式（7-68），得到 $\boldsymbol{F}_0$ 关于 $\boldsymbol{E}_0$ 的 PLS 回归方程

$$\boldsymbol{F}_0=\boldsymbol{E}_0w_1^*\boldsymbol{r}_1^{\mathrm{T}}+\boldsymbol{E}_0w_2^*\boldsymbol{r}_2^{\mathrm{T}}+\cdots+\boldsymbol{E}_0w_m^*\boldsymbol{r}_m^{\mathrm{T}}+\boldsymbol{F}_m \tag{7-69}$$

其中，$w_i^*=\prod_{k=1}^{i-1}(\boldsymbol{I}-w_kp_k')w_i(i=1,2,\cdots,m)$，$\boldsymbol{I}$ 为单位矩阵。

（3）还原为关于原始变量的 PLS 回归方程

将式（7-69）还原成关于原始变量的 PLS 回归方程

$$\hat{y}_k=\left(\overline{y}_k-\sum_{i=1}^{p}a_{ki}\frac{sy_k}{sx_i}\overline{x}_i\right)+\sum_{i=1}^{p}a_{ki}\frac{sy_k}{sx_i}x_i,\quad k=1,2,\cdots,q$$

其中，$\boldsymbol{a}_k$ 是矩阵 $\boldsymbol{a}_{p\times q}=\sum_{j=1}^{m}w_j^* r_j'$ 的第 k 个列向量；a_{ki} 是 $\boldsymbol{a}_k$ 的第 i 个分量。

3. 主成分提取

（1）第一轮主成分提取



求矩阵 $\boldsymbol{E}_0^{\mathrm{T}}\boldsymbol{F}_0\boldsymbol{F}_0^{\mathrm{T}}\boldsymbol{E}_0$ 的最大特征值所对应的单位特征向量 $\boldsymbol{w}_1$，得自变量的第一个主成分

$$\boldsymbol{t}_1=\boldsymbol{E}_0\boldsymbol{w}_1 \tag{7-70}$$

求矩阵 $\boldsymbol{E}_0^{\mathrm{T}}\boldsymbol{F}_0\boldsymbol{F}_0^{\mathrm{T}}\boldsymbol{E}_0$ 的最大特征值所对应的单位特征向量 $\boldsymbol{c}_1$，得因变量的第一个主成分

$$\boldsymbol{u}_1=\boldsymbol{F}_0\boldsymbol{c}_1 \tag{7-71}$$

求残差矩阵

$$\boldsymbol{E}_1=\boldsymbol{E}_0-\boldsymbol{t}_1\boldsymbol{p}_1^{\mathrm{T}} \tag{7-72}$$

$$\boldsymbol{F}_1=\boldsymbol{F}_0-\boldsymbol{t}_1\boldsymbol{r}_1^{\mathrm{T}} \tag{7-73}$$

式（7-72）中，$\boldsymbol{p}_1=\dfrac{\boldsymbol{E}_0^{\mathrm{T}}\boldsymbol{t}_1}{\|t_1\|^2}$；式（7-73）中，$\boldsymbol{r}_1=\dfrac{\boldsymbol{F}_0^{T}\boldsymbol{t}_1}{\|t_1\|^2}$。

在 PLS 方法中，称 $\boldsymbol{w}_1$ 为模型效应权重，$\boldsymbol{c}_1$ 为因变量权重，$\boldsymbol{p}_1$ 为模型效应载荷量。

（2）新一轮主成分提取

令 $\boldsymbol{E}_0=\boldsymbol{E}_1$，$\boldsymbol{F}_0=\boldsymbol{F}_1$，回到（1），对残差矩阵进行新一轮的主成分提取和回归分析。

设第 h 步的计算结果为

$$\boldsymbol{t}_h=\boldsymbol{E}_{h-1}\boldsymbol{w}_h \tag{7-74}$$

$$\boldsymbol{u}_h=\boldsymbol{F}_{h-1}\boldsymbol{c}_h \tag{7-75}$$

$$\boldsymbol{E}_h=\boldsymbol{E}_{h-1}-\boldsymbol{t}_h\boldsymbol{p}_h^{\mathrm{T}} \tag{7-76}$$

$$\boldsymbol{F}_h=\boldsymbol{F}_{h-1}-\boldsymbol{t}_h\boldsymbol{r}_h^{\mathrm{T}} \tag{7-77}$$

式（7-74）～式（7-77）中，$h=1,2,\cdots,m$，$m\leqslant \mathrm{rank}(E_0)$，$\boldsymbol{p}_h=\dfrac{\boldsymbol{E}_{h-1}^{\mathrm{T}}\boldsymbol{t}_h}{\|t_h\|^2}$，$\boldsymbol{r}_h=\dfrac{\boldsymbol{F}_{h-1}^{\mathrm{T}}\boldsymbol{t}_h}{\|t_h\|^2}$。

（3）主成分提取的终止准则

PLS 方法不需要选用所有的主成分建模，而是采用截尾的方法，即仅选择前 m 个主成分 $t_1,t_2,\cdots,t_m$，就可以得到一个预测性能较好的模型。因此，在主成分提取的每一轮计算中，都要对是否得到了足够多的主成分进行判断。

常用的判断准则有交叉有效性准则和复测定系数准则。

定义 7-1（交叉有效性）　称

$$Q_h^2=1-\frac{PRESS_h}{SS_{(h-1)}}$$

为主成分 t_h 关于因变量系统 Y 的交叉有效性。

上式中各参数的意义如下：$PRESS_h$ 是从所有 n 个样本点中舍弃某个样本点 $x^{(i)}(i=1,2,\cdots,n)$ 之后，用剩余的 $n-1$ 个样本点拟合出含 h 个主成分的回归方程，再对 $x^{(i)}(i=1,2,\ \cdots,n)$ 点

进行预测的预测误差平方和。更详细一些，记 $\hat{y}_{hj(-i)}$ 为 y_j 在样本点 $x^{(i)}$ 上的预测值，$PRESS_{hj}=\sum_{i=1}^{n}[y_{ij}-\hat{y}_{hj(-i)}]^2$ 为 y_j 的预测误差平方和，则 $PRESS_h=\sum_{i=1}^{p}PRESS_{hi}$ 就是 Y 的预测误差平方和。

$SS_{(h-1)}$ 是用所有 n 个样本点拟合出的含 $h-1$ 个主成分的回归方程的拟合误差平方和。更详细一些，记 $\hat{y}_{(h-1)ji}$ 为 y_j 在样本点 $x^{(i)}$ 上的拟合值，$SS_{(h-1)j}=\sum_{i=1}^{n}(y_{ij}-\hat{y}_{(h-1)ji})^2$ 为 y_j 的拟合误差平方和，则 $SS_{(h-1)j}=\sum_{i=1}^{p}SS_{(h-1)i}$ 就是 Y 的拟合误差平方和。

交叉有效性是对新增主成分能否对模型的预测功能有显著改进的判断指标。

若 $Q_h^2 \geqslant 1-0.95^2=0.0975$，则认为主成分 t_h 的边际贡献是显著的。

定义 7-2（复测定系数） 称

$$Q_h^2=\frac{\sum_{k=1}^{h}\left(\|t_k\|^2\times\|p_k\|^2\right)}{\|E_0\|^2}$$

为自变量系统 X 被提取的变异信息量。称

$$R_h^2=\frac{\sum_{k=1}^{h}\left(\|t_k\|^2\times\|r_k\|^2\right)}{\|F_0\|^2}$$

为回归方程的复测定系数。

复测定系数表示所提取的主成分的可解释变异信息占总变异信息的百分比。

当 $h=m$，复测定系数 R_m^2 的值足够大时，可在第 m 步终止主成分的提取计算。通常 $R_m^2 \geqslant 0.85$ 即可。

7.3.3 偏最小二乘回归方法的辅助分析

PLS 方法除了前面讲的建模技术，还包括 PLS 辅助分析技术，可以在获得一个更为合理的回归模型的同时，完成一些类似于主成分分析和典型相关分析的研究内容，提供更加丰富、深入的系统信息。

1. 自变量和因变量之间的相关关系分析

在一元回归分析中，为了判定自变量和因变量之间的关系，经常采用散点图来作直观的分析，简单而有效。这种方法在多元回归分析中遇到困难：多维数据构成了一个超平面，难以作直观观察；各自变量间相互关联，不能将变量简单地分割开来分析。

PLS 方法的 t_1/u_1 平面图功能使这一点成为可能。

在 PLS 方法中，自变量集合 X 和因变量集合 Y 之间的相关关系可以通过 t_1 和 u_1 的相关关系得到反映。因此，绘制以 t_1 为横坐标，u_1 为纵坐标的 t_1/u_1 平面图，绘出第一主成分偶

对 (t_1,u_1) 的观测样本散点图。如果所有样本点 $(t_1(i),u_1(i))$ $(i=1,2,\cdots,n)$ 在图中的排列近似于一条直线，则说明 X 和 Y 之间存在着较强的相关关系，这时采用 PLS 方法建立 Y 对 X 的线性模型才会是合理的。



2. 主成分对变量的解释能力的评价

在 PLS 计算过程中，要求所提取的自变量主成分 t_h 尽可能多地代表 X 的变异信息，尽可能与 Y 相关联，解释 Y 中的信息。为了测量 t_h 对 X 和 Y 的解释能力，现给出如下定义。

定义 7-3（自变量的主成分对自变量系统的各种解释能力） ① 称主成分 t_h 与自变量 x_j 的简单相关系数的平方

$$\mathrm{Rd}(x_j;t_h)=r^2(x_j;t_h)$$

为 t_h 对某个自变量 x_j 的解释能力。

② 称

$$\mathrm{Rd}(X;t_h)=\frac{1}{p}\sum_{j=1}^{p}\mathrm{Rd}(x_j;t_h)$$

为 t_h 对自变量系统 X 的解释能力。

③ 称

$$\mathrm{Rd}(x_j;t_1,t_2,\cdots,t_m)=\sum_{h=1}^{m}\mathrm{Rd}(x_j;t_h)$$

为 $t_1,t_2,\cdots,t_m$ 对某个自变量 x_j 的累计解释能力。

④ 称

$$\mathrm{Rd}(X;t_1,t_2,\cdots,t_m)=\sum_{h=1}^{m}\mathrm{Rd}(X;t_h)$$

为 $t_1,t_2,\cdots,t_m$ 对自变量系统 X 的累计解释能力。

定义 7-4（自变量的主成分对因变量系统的各种解释能力） ① 称主成分 t_h 与因变量 y_j 的简单相关系数的平方

$$\mathrm{Rd}(y_j;t_h)=r^2(y_j;t_h)$$

为 t_h 对某个因变量 y_j 的解释能力。

② 称

$$\mathrm{Rd}(Y;t_h)=\frac{1}{q}\sum_{j=1}^{q}\mathrm{Rd}(y_j;t_h)$$

为 t_h 对因变量系统 Y 的解释能力。

③ 称

$$\mathrm{Rd}(y_j;t_1,t_2,\cdots,t_m)=\sum_{h=1}^{m}\mathrm{Rd}(y_j;t_h)$$

为 $t_1,t_2,\cdots,t_m$ 对某个自变量 y_j 的累计解释能力。

④ 称

$$\mathrm{Rd}(Y;t_1,t_2,\cdots,t_m)=\sum_{h=1}^{m}\mathrm{Rd}(Y;t_h)$$

为 $t_1,t_2,\cdots,t_m$ 对因变量系统 Y 的累计解释能力。

3. 自变量对因变量系统的解释能力

在 PLS 方法中，自变量对因变量的解释能力是以变量投影重要性指标（VIP）来测度的。

定义 7-5（自变量对主成分的边际贡献）　称

$$\mathrm{VIP}_j=\sqrt{\frac{p}{\mathrm{Rd}(Y;t_1,t_2,\cdots,t_m)}\sum_{h=1}^{m}\mathrm{Rd}(Y;t_h)w_{hj}^2}$$

为自变量 x_j 对主成分 t_h 的边际贡献。其中，w_{hj} 是主轴 w_h 的第 j 个分量；$\mathrm{Rd}(Y;t_h)$，$\mathrm{Rd}(Y;t_1,t_2,\cdots,t_m)$ 分别是 t_h 对 Y 的解释能力和 $t_1,t_2,\cdots,t_m$ 对 Y 的累计解释能力。

VIP_j 定义式的意义基于这样一个事实：由于 x_j 对 Y 的解释是通过 t_h 来传递的，如果 t_h 对 Y 的解释能力很强，而 x_j 在构造 t_h 时又起到了相当重要的作用，则 x_j 对 Y 的解释能力就被视为很强。也就是说，如果在 $\mathrm{Rd}(Y;t_h)$ 值很大时的成分 t_h 上，w_{hj} 取得很大的值，则 x_j 对解释 Y 就有很重要的作用。

另外，容易证明 $\sum_{j=1}^{p}\mathrm{VIP}_j^2=p$，所以，对于 p 个自变量 $x_j\ (j=1,2,\cdots,p)$，如果它们在解释 Y 时的作用都相同，则所有 VIP_j 均等于 1；否则，对于 VIP_j（$\mathrm{VIP}_j>1$）很大的 x_j，它在解释因变量 Y 时就有更加重要的作用。

统计工具箱提供了两个主成分分析函数 princomp 和 pcacov。

（1）princomp 函数

其调用格式如下：

```
[COEFF,SCORE] = princomp(X)
[COEFF,SCORE,latent] = princomp(X)
[COEFF,SCORE,latent,tsquare] = princomp(X)
[...] = princomp(X,'econ')
```

其中，$\boldsymbol{X}$ 是 $n\times p$ 的原始数据矩阵；COEFF 为返回主成分的系数，为 p 阶矩阵，每一列为一个主成分的系数；SCORE 为返回原数据在新坐标系中的新数据；latent 返回协方差矩阵 $\boldsymbol{X}$ 的特征值；tsquare 返回每个数据点的 Hotelling 统计量。

（2）pcacov 函数

其调用格式如下：

```
COEFF = pcacov(V)
[COEFF,latent] = pcacov(V)
```

其中，$\boldsymbol{V}$ 是协方差矩阵；COEFF 为返回主成分；latent 为返回协方差矩阵 $\boldsymbol{V}$ 的特征值。

统计工具箱自带了数据 cities.mat，它是反映美国 329 个城市生活水平的 9 个不同的指标

数据。这 9 个指标包括气候、住房、健康、犯罪率、交通、教育、艺术、娱乐及经济状态。对每一个指标，值越高越好。下面通过主成分分析减少变量的数目。

【例 7-8】 反映城市生活水平的不同指标的主成分分析。

其实现的 MATLAB 程序代码如下：



```
>> clear all;
load cities;    %载入原始数据
%标准化数据
stdr=std(ratings);
sr=ratings./repmat(stdr,329,1);
%第一种主成分分析方法
[pcs,newdata,variances,t2]=princomp(sr);
pcs
plot(newdata(:,1),newdata(:,2),'*');
xlabel('第一个主成分');ylabel('第二个主成分');
```

运行程序，输出如下：

```
pcs =
0.2064    0.2178   -0.6900    0.1373   -0.3691    0.3746   -0.0847   -0.3623    0.0014
0.3565    0.2506   -0.2082    0.5118    0.2335   -0.1416   -0.2306    0.6139    0.0136
0.4602   -0.2995   -0.0073    0.0147   -0.1032   -0.3738    0.0139   -0.1857   -0.7164
0.2813    0.3553    0.1851   -0.5391   -0.5239    0.0809    0.0186    0.4300   -0.0586
0.3512   -0.1796    0.1464   -0.3029    0.4043    0.4676   -0.5834   -0.0936    0.0036
0.2753   -0.4834    0.2297    0.3354   -0.2088    0.5022    0.4262    0.1887    0.1108
0.4631   -0.1948   -0.0265   -0.1011   -0.1051   -0.4619   -0.0215   -0.2040    0.6858
0.3279    0.3845   -0.0509   -0.1898    0.5295    0.0899    0.6279   -0.1506   -0.0255
0.1354    0.4713    0.6073    0.4218   -0.1596    0.0326   -0.1497   -0.4048    0.0004
```

具体地，以第一列为例（即第一个主成分），最大的权值是第三个和第七个元素，分别对应于变量“健康”和“艺术”。

第二个输出为原始数据在主成分定义的坐标系中的新数据，其矩阵大小与原始数据相同。图 7-12 显示的是原始数据在前面两个主成分上的投影。

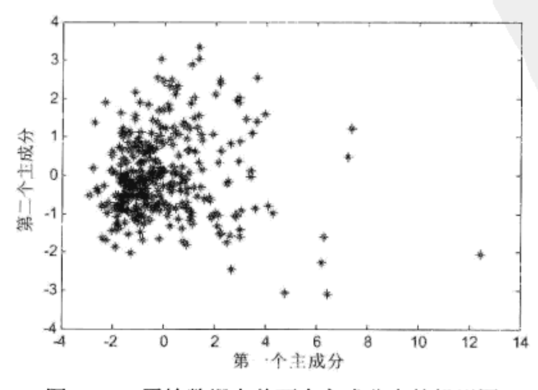


图 7-12 原始数据在前两个主成分上的投影图

第三个输出为协方差矩阵的特征值

```
>> variances'
ans =
  3.4083    1.2140    1.1415    0.9209    0.7533    0.6306    0.4930    0.3180    0.1204
```

由累计特征值可知，前 5 个主成分占了总方差的 82.7%，因此只需要 5 个变量（气候、住房、健康、犯罪率、交通）就可以表征不同城市的生活水平。

【例 7-9】 根据表 7-10 中人体头发的元素分析结果进行主成分分析。

表 7-10 人体头发的元素分析

样　本	Cu	Mn	Cl	Br	I
1	9.2	0.30	1770	12.0	3.6
2	12.4	0.39	930	50.0	2.3
3	7.2	0.32	2750	65.3	3.4
4	10.2	0.36	1500	3.4	5.3
5	10.1	0.50	1040	39.2	1.9
6	6.5	0.20	2490	90.0	4.6
7	5.6	0.29	2940	88.0	5.6
8	11.8	0.42	867	43.1	1.5
9	8.5	0.25	1620	5.2	6.2

其实现的 MATLAB 程序代码如下：

```
>> clear all;
x=[9.2,0.30,1770,12.0,3.6;12.4,0.39,930,50.0,2.3;7.2,0.32,2750,65.3,3.4;...
    10.2,0.36,1500,3.4,5.3;10.1,0.50,1040,39.2,1.9;6.5,0.20,2490,90.0,4.6;...
    5.6,0.29,2940,88.0,5.6;11.8,0.42,867,43.1,1.5;8.5,0.25,1620,5.2,6.2];
stdr=std(x);
sr=x./stdr(ones(9,1),:);
[pcs,newdata,variances,t2]=princomp(sr);   %主成分分析
pcs             %主成分
newdata         %得分
variances       %方差
t2              %统计量
plot(newdata(:,1),newdata(:,2),'*');
gname           %获取各点代表的样本
```

运行程序，输出如下（主成分得分图见图 7-13）

```
pcs =
   -0.5215    0.1028   -0.4127    0.1820   -0.7170
   -0.4652   -0.2691    0.7899    0.2833   -0.0829
    0.5174   -0.1704    0.3127   -0.3823   -0.6778
    0.2769   -0.7610   -0.2824    0.5140   -0.0175
    0.4090    0.5558    0.1680    0.6901   -0.1393
```



```
newdata =
   -0.1658    0.7783   -0.0895   -0.6913    0.0216
   -1.8837   -0.4626   -0.6649    0.3083   -0.2388
    1.2205   -0.8723    0.3504   -0.5121   -0.2328
   -0.5357    1.4566    0.3854    0.2579   -0.2436
   -2.0422   -0.7931    0.7886    0.0859    0.3139
    2.3119   -0.6715   -0.7569    0.0573    0.2033
    2.5738   -0.7007    0.4732    0.4217   -0.0676
   -2.1928   -0.6658   -0.3444   -0.0465    0.0418
    0.7139    1.9312   -0.1420    0.1188    0.2022
variances =
    3.3513
    1.1807
    0.2849
    0.1383
    0.0448
t2 =
    4.0149
    4.7531
    4.6267
    4.2103
    6.2156
    4.9348
    4.5668
    2.2815
    4.3964
```

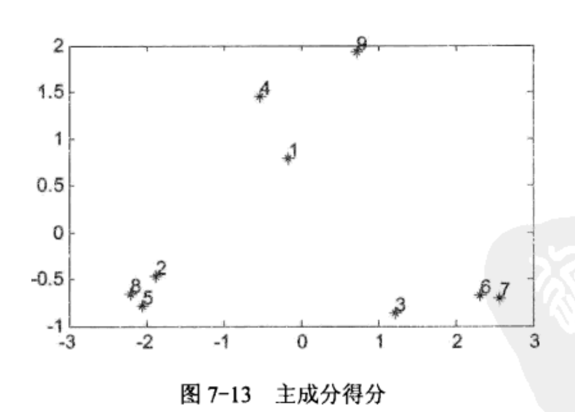


图 7-13 主成分得分

从变量（variances）结果可以看出，共有 5 个主成分，但前面两个的主成分作用显著，占了总方差的约 91%。

第 8 章　多元统计分析

多元素分析是数据统计学中近三四十年来迅速发展的重要分支之一。由于计算机及软件的使用日益广泛，多元统计分析的方法已在生物、医学、地质、农业、工程技术、气象和社会经济等许多学科，得到日益广泛的应用。

8.1　引言

在日常生活和科学研究过程中，往往同时观测 n 个对象的 p 个属性，然后再对这些数据进行整理分析，从而得出所期望的结论。多元统计分析就是处理这类问题的一个有力工具。

如果同时研究一个总体的 p 个属性，则可以把这个总体看成一个 p 元向量。从总体中随机抽取进行观测的对象叫做样本，样本的一次观测结果有 p 个数值，可以看做是这个 p 元向量的一次取值。第 i 个样本的第 j 个属性的观测结果记为 x_{ij}，每个样本可以用一个 p 维向量来表示

$$\boldsymbol{x}_i=\begin{pmatrix}x_{1i}\\x_{2i}\\\vdots\\x_{pi}\end{pmatrix}$$

对 n 个样本进行观测的全部结果，共有 $p\times n$ 个数据，可以用下列矩阵表示

$$\boldsymbol{X}=(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n)=\begin{pmatrix}x_{11}&x_{12}&\cdots&x_{1n}\\x_{21}&x_{22}&\cdots&x_{2n}\\\vdots&\vdots&&\vdots\\x_{p1}&x_{p2}&\cdots&x_{pn}\end{pmatrix}$$

一个多元总体可以看成一个多元随机变量。实际中，考察一个 p 元总体就是考察这个总体中每个对象的 p 个属性，或者说考察一个 p 元随机变量。多元统计分析的主要任务包括分析各观测数据之间的关系，以及推断总体的某些性质。

同一元样本的数字特征一样，也可以定义多元样本的数字特征。

（1）样本的平均值

$$\bar{\boldsymbol{X}}=\begin{pmatrix}\bar{\boldsymbol{x}}_1\\\vdots\\\bar{\boldsymbol{x}}_p\end{pmatrix}=\begin{pmatrix}\sum_{i=1}^{n}x_{1i}/n\\\vdots\\\sum_{i=1}^{n}x_{pi}/n\end{pmatrix}$$

样本的平均值就是各变量的样本平均值组成的向量，它是 n 个样本的重心。

（2）中心化数据

常常需要将原始数据减去它的均值，称为中心化数据 $\underline{\boldsymbol{X}}$ 。



$$\underline{\boldsymbol{X}}=\begin{pmatrix} x_{11}-\overline{x}_1 & x_{12}-\overline{x}_1 & \cdots & x_{1n}-\overline{x}_1 \\ x_{21}-\overline{x}_2 & x_{22}-\overline{x}_2 & \cdots & x_{2n}-\overline{x}_2 \\ \vdots & \vdots & & \vdots \\ x_{p1}-\overline{x}_p & x_{p2}-\overline{x}_p & \cdots & x_{pn}-\overline{x}_p \end{pmatrix}$$

（3）标准化数据

$$\tilde{\boldsymbol{X}}=\begin{pmatrix} \dfrac{x_{11}-\overline{x}_1}{s_1} & \dfrac{x_{12}-\overline{x}_1}{s_1} & \cdots & \dfrac{x_{1n}-\overline{x}_1}{s_1} \\ \dfrac{x_{21}-\overline{x}_2}{s_2} & \dfrac{x_{22}-\overline{x}_2}{s_2} & \cdots & \dfrac{x_{2n}-\overline{x}_2}{s_2} \\ \vdots & \vdots & & \vdots \\ \dfrac{x_{p1}-\overline{x}_p}{s_p} & \dfrac{x_{p2}-\overline{x}_p}{s_p} & \cdots & \dfrac{x_{pn}-\overline{x}_p}{s_p} \end{pmatrix}$$

（4）距离

距离是数学中的一个抽象概念，它可以用于描述样本之间的差异程度。常用的距离有欧氏距离、马氏距离和绝对距离。

欧氏距离的定义为

$$d_{ij}=\sum_{k=1}^{n}(x_{ki}-x_{kj})^2=(x_i-x_j)(x_i-x_j)$$

马氏距离的定义为

$$d_{ij}=(x_i-x_j)\boldsymbol{S}^{-1}(x_i-x_j)$$

式中，$\boldsymbol{S}$ 是协方差矩阵。

绝对距离的定义为

$$d_{ij}=\sum_{k=1}^{n}\left|x_{ki}-x_{kj}\right|$$

8.2 因素分析

多元数据常常包含大量的测量变量，有时候这些变量是相互重叠的。也就是说，它们之间存在相关性。因素分析的概念是英美心理统计学者们最早提出的，因素分析法的目的就是从试验所得的 $m\times n$ 个数据样本中概括和提取出较少量的关键因素，它们能反映和解释所得的大量观测事实，从而建立起最简洁、最基本的概念系统，揭示出事物之间最本质的联系。

8.2.1 因素分析的理论介绍

因素分析的数学模型如下：

$$Y = Pf + s \tag{8-1}$$

式中，$\boldsymbol{Y}=[\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_m]^{\mathrm{T}}$ 为可观测的 m 维随机向量；任一分量 $\boldsymbol{y}_i$ 是一随机时间序列变量，记作 $\boldsymbol{y}_i=(y_{1i},y_{2i},\cdots,y_{ki})^{\mathrm{T}}$；$y_i$ 称为公共因素向量 $(K\leqslant i)$；$\boldsymbol{s}=(s_1,s_2,\cdots,s_m)^{\mathrm{T}}$ 为特殊因素向量；$\boldsymbol{P}$ 为因素负荷矩阵 $(m\times q)$；$\boldsymbol{f}$，$\boldsymbol{s}$ 都是相互无关的随机向量，一般是不可观测的。

为了计算方便，经常将随机向量 $\boldsymbol{Y}$ 进行标准化。假设进行了 n 次观测，标准化记作 $\boldsymbol{Z}$，且 $\boldsymbol{Z}=[\boldsymbol{z}_1,\boldsymbol{z}_2,\cdots,\boldsymbol{z}_m]^{\mathrm{T}}$，其中第 i 个分量第 j 次测定的标准值为

$$z_{ij}=\frac{y_{ij}-\mu_i}{\sigma_i^2},\quad i=1,2,\cdots,m;\ j=1,2,\cdots,n$$

其中，$\mu_i=\sum_{j=1}^{n}x_{ij}/n$ 是第 i 个变量的观测均值，σ_i^2 是第 i 个变量的观测方差。这样，因素分析的模型可以重新写成

$$Z = Pf + s \tag{8-2}$$

具体展开为

$$z_{ij}=\sum_{k=1}^{q}p_{ik}f_{kj}+s_{ij} \tag{8-3}$$

上式的意义表示第 i 个分量第 j 次测定标准值与公共因素、特殊因素的关系。因素负荷矩阵的统计意义是：$\boldsymbol{P}$ 的行元素的平方和代表公共因素对变量 z_i 的方差所作的贡献，称为共性方差，它的大小反映了变量 z_i 对公共因素的依赖限度；$\boldsymbol{P}$ 的列元素的平方和代表第 k 个公共因素 f_k 对向量 $\boldsymbol{Z}$ 的影响，称为方差贡献，它的大小反映了随机向量 $\boldsymbol{Z}$ 对 f_k 的依赖程度，是衡量公共因素 f_k 相对重要性的一个重要尺度。

因素分析模型的物理意义解释如下：假设将每个因素看成一个坐标轴，q 个因素变量构成了一个 q 维的因素空间，式（8-1）和式（8-2）就是将原来的 m 个观测变量投影到 q 维的因素空间，用 q 个因素变量的组合来表达原观测变量的主要信息甚至全部信息。

随机向量 $\boldsymbol{Z}$ 的协方差矩阵与负荷矩阵的关系如下

$$\mathrm{Cov}(\boldsymbol{Z},\boldsymbol{Z})=\boldsymbol{P}\boldsymbol{P}^{\mathrm{T}}+\boldsymbol{\Phi} \tag{8-4}$$

因素分析是从一组向量的相互关系出发，建立若干个相互正交的因素轴，将这组向量最大限度地包含在因素空间内，使各个向量在各因素轴上的投影和达到最大。因素分析的关键是从变量的相关矩阵中，利用式（8-4）求解出因素负荷矩阵 $\boldsymbol{P}$。

8.2.2 因素分析的函数介绍

因素分析一般有两步：第一步是从信号的相关矩阵 $\boldsymbol{R}$ 中求解出无限多个 $\boldsymbol{P}$ 中的一个，确定因素数目，称为因素提取过程；第二步是经过旋转变换，找到一个最合适的 $\boldsymbol{P}$，称为因素旋转过程。通过因素提取过程得到了若干个因素之后，因素的含义往往不明确，为了对因

素作出解释，就需要对因素负荷矩阵进行旋转变换。

统计工具箱中提供了因素负荷矩阵的极大似然估计函数 factoran。

其调用格式如下：



```
lambda = factoran(X,m)
[lambda,psi] = factoran(X,m)
[lambda,psi,T] = factoran(X,m)
[lambda,psi,T,stats] = factoran(X,m)
[lambda,psi,T,stats,F] = factoran(X,m)
[...] = factoran(...,param1,val1,param2,val2,...)
```

其中，*X* 是观测向量；*m* 是公共因素的数目；'param1'，value1 等是控制模型和输出的名称/数值对（可选参数）；lambda 返回因素负荷矩阵的估计值；psi 返回特殊因素负荷矩阵的估计值；*T* 返回因素负荷旋转矩阵；stats 是一个数据结构，它包含了与假设检验有关的信息。

【例 8-1】 对 460 种不同汽车的 5 项指标数据进行两因素分析（其中，carbig 数据是 MATLAB 统计工具箱自带的）。

其实现的 MATLAB 程序代码如下：

```
>> clear all;
load carbig
X = [Acceleration Displacement Horsepower MPG Weight];
X = X(all(~isnan(X),2),:);
%估计因素负荷矩阵
[Lambda,Psi,T,stats,F] = factoran(X,2,'scores','regression');
Lambda      %输出因素负荷矩阵
inv(T'*T) % F 的相关矩阵
Lambda*Lambda'+diag(Psi) % X 的相关矩阵
Lambda*inv(T)                  % 未经旋转的因素负荷矩阵
F*T';                          % 未经旋转的因素贡献率
%绘制未经旋转的负荷点和旋转斜坐标
invT=inv(T);
Lambda0=Lambda*invT;
biplot(Lambda,'LineWidth',2,'MarkerSize',20);
line([-invT(1,1),invT(1,1),NaN,-invT(2,1),invT(2,1)],...
     [-invT(1,2),invT(1,2),NaN,-invT(2,2),invT(2,2)]);
xlabel('载入因素 1');
ylabel('载入因素 2');
```

运行程序，输出如下：

公共因素负荷矩阵为

```
Lambda =
   -0.2432   -0.8500
    0.8773    0.3871
    0.7618    0.5930
   -0.7978   -0.2786
    0.9692    0.2129
```

可见，第二、第三和第五个指标与第一个因素有关。

因素相关矩阵为

```
ans =
    1.0000   -0.0000
   -0.0000    1.0000
```

5 项指标之间的相关矩阵为

```
ans =
    1.0000   -0.5424   -0.6893    0.4309   -0.4167
   -0.5424    1.0000    0.8979   -0.8078    0.9328
   -0.6893    0.8979    1.0000   -0.7730    0.8647
    0.4309   -0.8078   -0.7730    1.0000   -0.8326
   -0.4167    0.9328    0.8647   -0.8326    1.0000
```

未经旋转的负荷矩阵为

```
ans =
   -0.5020    0.7277
    0.9550   -0.0865
    0.9113   -0.3185
   -0.8450    0.0091
    0.9865    0.1079
```

未经旋转的因素负荷点及旋转斜轴如图 8-1 所示。

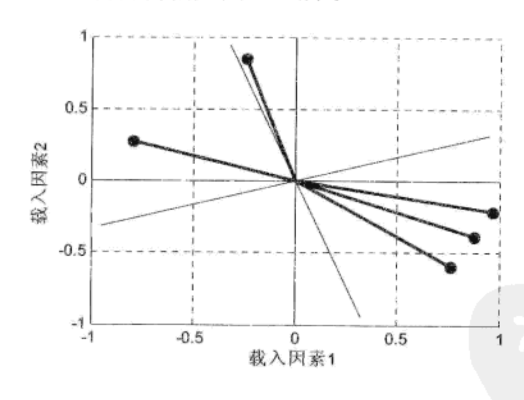


图 8-1　未经旋转的因素负荷点位置

8.2.3 因素分析的应用示例分析

影响股票价格的因素分析。为此，记录了 100 周的时间内，10 家公司的股票价格的变化。在这 10 家公司中，4 家公司属于一般的技术公司，3 家公司属于金融公司，3 家公司属于零售公司。从原理上说，同一类型公司的股票价格应该同时变化，下面通过因素分析对此进行定量分析，这里的因素就是公司的类型。

【例 8-2】 影响股票价格的因素分析（此数据是 MATLAB 自带的）。



其实现的 MATLAB 程序代码如下：

```
>> clear all;
load stockreturns              %装载数据
m=3;                           %因素个数
%因素分析
[loadings,specifivVar,T,stats]=factoran(stocks,m,'rotate','none');
loadings                       %未经旋转的公共因素负荷矩阵
specifivVar                    %未经旋转的特殊因素矩阵
%因素分析
[loadingsPM,specifivVarPM]=factoran(stocks,m,'rotate','promax');
loadingsPM                     %旋转后的公共因素负荷矩阵
figure;
subplot(121);
plot(loadingsPM(:,1),loadingsPM(:,2),'r.');
text(loadingsPM(:,1),loadingsPM(:,2),num2str((1:10)'));
line([-1 1 NaN 0 0 NaN 0 0],[0 0 NaN -1 1 NaN 0 0],'color','red');
xlabel('因素 1');ylabel('因素 2');
axis square;
subplot(122);
plot(loadingsPM(:,1),loadingsPM(:,3),'r.');
text(loadingsPM(:,1),loadingsPM(:,3),num2str((1:10)'));
line([-1 1 NaN 0 0 NaN 0 0],[0 0 NaN -1 1 NaN 0 0],'color','red');
xlabel('因素 1');ylabel('因素 3');
axis square;
```

运行程序，输出如下：

未经旋转的公共因素负荷矩阵为

```
loadings =
    0.8885    0.2367   -0.2354
    0.7126    0.3862    0.0034
    0.3351    0.2784   -0.0211
    0.3088    0.1113   -0.1905
    0.6277   -0.6643    0.1478
    0.4726   -0.6383    0.0133
    0.1133   -0.5416    0.0322
    0.6403    0.1669    0.4960
    0.2363    0.5293    0.5770
    0.1105    0.1680    0.5524
```

从上述公共因素负荷矩阵可知，难以与已知的 3 种类型的公司相对应，原因在于未经旋转的因素负荷矩阵难以解释。

特殊因素矩阵为

```
> specifivVar    %未经旋转的特殊因素矩阵
specifivVar = 0.0991
```

```
0.3431
0.8097
0.8559
0.1429
0.3691
0.6928
0.3162
0.3311
0.6544
```

由特殊因素矩阵可以看出，股票价格的变化还受到某种特殊因素的影响。

旋转后的公共因素负荷矩阵为

```
loadingsPM =
    0.9452    0.1214   -0.0617
    0.7064   -0.0178    0.2058
    0.3885   -0.0994    0.0975
    0.4162   -0.0148   -0.1298
    0.1021    0.9019    0.0768
    0.0873    0.7709   -0.0821
   -0.1616    0.5320   -0.0888
    0.2169    0.2844    0.6635
    0.0016   -0.1881    0.7849
   -0.2289    0.0636    0.6475
```

由上述数据明显可以看出，第一~第四家公司属于同一类，与第一个因素有关；第五~第七家公司属于同一类，与第二个因素有关；第八~第十家公司属于同一类，与第三个因素有关。

在上述因素旋转的过程中，采用的是斜交旋转（promax 准则）。这种旋转方式在负荷中产生一个简单的结构，即大多数的股票价格仅仅对一个因素有较大的负荷。为了看清楚这种结构，可以使用因素负荷为坐标绘制负荷矩阵，如图 8-2 所示。

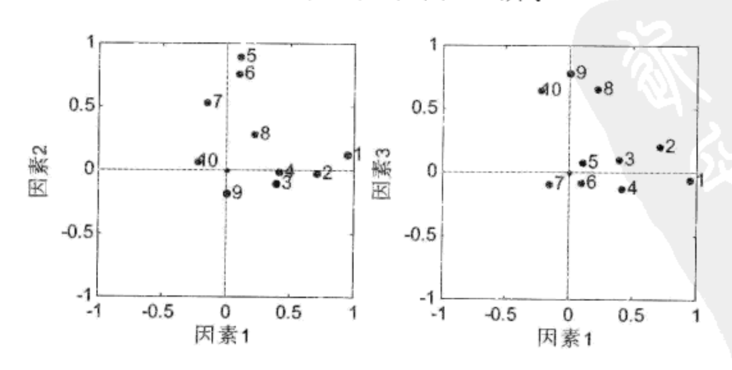


图 8-2　斜交旋转后的负荷矩阵结构

由图 8-2 可以看出，第一个因素轴对应金融公司，第二个因素轴对应零售公司，第三个因素轴对应一般的技术公司。

8.3 聚类分析



人类认识世界的一种重要方法是将世界上的事物进行分类，从中发现规律，进而改造世界。正因为这样，分类学早就成为人类认识世界的一门基础学科。由于事物的复杂性，单凭经验来分类是远远不够的，利用数学方法进行更科学的分类成为一种必然的趋势。随着计算机的普及，利用数学方法研究分类不仅非常必要，而且完全可能。因此，聚类分析作为多元分析的一个重要分支，发展非常迅速。

8.3.1 聚类分析的理论介绍

在分类学中，一般把某种性质比较相近的事件归为同一类，把性质不相近的事件归为不同的类。利用数学方法的分类是建立在各个事物关于其性质变量的测量数据基础上的，即利用这些数据的内在联系和规律来进行分类。为此，首先需要有刻画各个变量之间或各个事物之间关系密切程度的描述。目前，描述变量之间关系的数学方法很多，常用的是相似（或相关）系数和距离。

1. 相似系数

假设测定了n个变量$x_1,x_2,\cdots,x_n$的M组数据，记作

$$x_{1m},x_{2m},\cdots,x_{nm}\,,\quad m=1,2,\cdots,M$$

这样，n个变量就可以看做是$\Re^k$空间中的n个向量，则向量x_i，x_j之间的相关性，即相关系数可以定义如下：

$$r_{ij}=\frac{\sum_{m=1}^{M}(x_{mi}-\overline{x}_i)(x_{mj}-\overline{x}_j)}{\sqrt{\sum_{m=1}^{M}(x_{mi}-\overline{x}_i)^2(x_{mj}-\overline{x}_j)^2}}$$

式中，$x_i=\dfrac{1}{M}\sum_{m=1}^{M}x_{mi}$；$x_j=\dfrac{1}{M}\sum_{m=1}^{M}x_{mj}$。

当然，除了上述定义之外，还有其他的相关系数的定义，读者可以参考相关书籍。相关系数（或相似系数）具有以下性质：

1）$|r_{ij}|\leqslant 1,\forall i,j$。

2）$r_{ij}=r_{ji},\forall i,j$。

而且$|r_{ij}|$越接近于 1，说明x_i，x_j越相似或相关；$|r_{ij}|$越接近于 0，说明x_i，x_j越不相似或不相关。特别地，$|r_{ij}|=1$时，说明$x_i=ax_j$，即x_i，x_j是完全线性相关的；$|r_{ij}|=0$时，说明x_i，x_j是正交的。

2. 距离

在欧氏空间中，两个向量x_i，x_j除了用它们的夹角的余弦来度量它们的相似程度外，

还可用它们的距离来度量。常用的距离有以下几种：

（1）欧氏距离

$$d_{ij}=\sqrt{\sum_{m=1}^{M}(x_{mi}-x_{mj})^2}$$

（2）Minkowski 距离

$$d_{ij}=\left(\sum_{m=1}^{M}\left|x_{mi}-x_{mj}\right|^p\right)^{1/p}$$

式中，p 是一正整数。当 $p=2$ 时，即为欧氏距离。当 $p=1$ 时，有

$$d_{ij}=\sum_{m=1}^{M}\left|x_{mi}-x_{mj}\right|$$

称为绝对值距离。

（3）切比雪夫距离

$$d_{ij}=\max_{1\leqslant m\leqslant M}\left|x_{mi}-x_{mj}\right|$$

8.3.2 聚类分析的函数介绍

统计工具箱实现了两类聚类方法，即系统聚类法和 K-均值聚类法。

（1）系统聚类法

系统聚类法是目前用得最多的一种聚类方法。它的基本思想是：首先，将要分类的 n 个变量各自看做一类，然后计算各类之间的关系密切程度（相关系数或距离），并将关系最密切的两类归为一类，其余不变，即得到 $n-1$ 个类，如此重复进行下去，每次归类都减少一类，直至最后，n 个变量都归为一类。这一归类过程可以用一张聚类图形象地表示出来，由聚类图明显可以看出分类过程。

统计工具箱实现系统聚类法的基本步骤如下：

① 计算数据集每对元素之间的距离，对应函数为 pdist。

其调用格式如下：

```
y = pdist(X)
y = pdist(X,metric)
y = pdist(X,distfun)
y = pdist(X,'minkowski',p)
```

其中，X 是 $m\times n$ 的矩阵，表示 m 个大小为 n 的向量；metric 是计算距离的方法选项，其选项含义如下：distfun 是自定义的距离函数；p 是自定义距离函数的输入参数；y 返回大小为 $m(m-1)/2$ 的距离矩阵，距离的排列顺序为(1, 2)，(1, 3),⋯, (1, m), (2, 1),⋯,(2, m),⋯, (m-1,m)，y 也称为相似矩阵。

- metric=euclidean 时：表示欧氏距离（默认值）。
- metric=seuclidean 时：表示标准的欧氏距离。
- metric=mahalanbis 时：表示 mahalanbis 距离。



② 对变量进行分类，构成一个系统聚类树，对应函数为 linkage。

其调用格式如下：

```
Z = linkage(y)
Z = linkage(y,method)
```

其中，y 是距离向量；Z 为返回系统聚类树；method 是采用的算法选项，其取值如下：

- method=single 时：表示最短距离。
- method=complete 时：表示最长距离。
- method=average 时：表示平均距离。
- method=centroid 时：表示中心距离。

③ 确定怎样划分系统聚类树，得到不同的类，对应的函数为 cluster。

其调用格式如下：

```
T = cluster(Z,'cutoff',c)
T = cluster(Z,'cutoff',c,'depth',d)
T = cluster(Z,'cutoff',c,'criterion',criterion)
T = cluster(Z,'maxclust',n)
```

其中，Z 是系统聚类树，为 $(m-1)\times 3$ 的矩阵；c 是阈值；n 是类的最大数目；criterion 是聚类的准则；d 是树的深度；T 是一个大小为 m 的向量，它包括原始数据每个观测量的编号；maxclust 为聚类的选项；depth 指定系统聚类树的水平数，并包含在不连续系数的计算中；cutoff 是一个临界值，它决定 cluster 函数怎样聚类。

【例 8-3】 利用系统聚类法对以下 5 个变量分类。

其实现的 MATLAB 程序代码如下：

```
>> clear all;
X=[1 2;2.5 4.5;2 2;4 1.5;4 2.5];   %分析数据矩阵
%显示 5 个变量的位置
figure(1);
plot(X(:,1),X(:,2),'*');
grid on;axis([0 5 0 5]);gname
%计算变量之间的距离信息
Y=pdist(X);
DisM=squareform(Y)
Z=linkage(Y)  %生成系统聚类树
%显示系统聚类树
figure(2);dendrogram(Z);
%不同阈值的分类结果
T1=cluster(Z,2)
T2=cluster(Z,3)
T3=cluster(Z,5)
```

运行程序，输出如下：

5 个变量在空间的位置如图 8-3 所示。

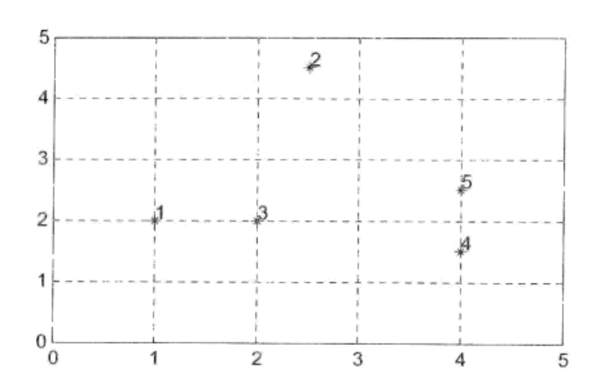


图 8-3　5 个变量在空间的位置

各个变量之间的距离矩阵为

```
DisM =
            0    2.9155    1.0000    3.0414    3.0414
       2.9155         0    2.5495    3.3541    2.5000
       1.0000    2.5495         0    2.0616    2.0616
       3.0414    3.3541    2.0616         0    1.0000
       3.0414    2.5000    2.0616    1.0000         0
```

系统聚类树连接信息矩阵为

```
Z =
       4.0000    5.0000    1.0000
       1.0000    3.0000    1.0000
       6.0000    7.0000    2.0616
       2.0000    8.0000    2.5000
```

系统聚类树图如图 8-4 所示。

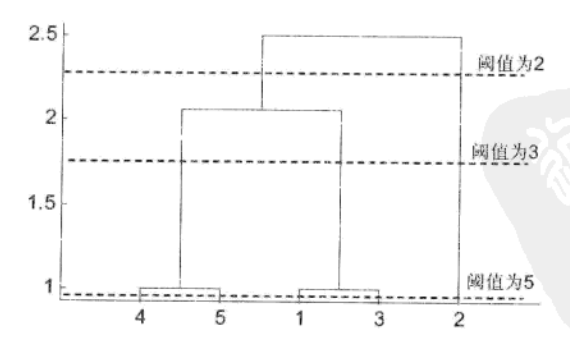


图 8-4　系统聚类树图

当阈值为 2 时的聚类结果为

```
T1 =
       2    1    2    2    2
```

即这 5 个变量分为两类——{1, 3, 4, 5}，{2}。

当阈值为 3 时的聚类结果为

```
T2 =
     2     3     2     1     1
```

即这 5 个变量分为 3 类——{1, 3}，{2}，{4, 5}。

当阈值为 5 时的聚类结果为

```
T3 =
     1     2     3     4     5
```

即这 5 个变量分为 5 类——{1}，{2}，{3}，{4}，{5}。

（2）K-均值聚类法

K-均值聚类法是一种简单、高效的聚类算法。假设有 n 个变量 $x_1,x_2,\cdots,x_n$，现将 n 个变量划分为 K 个类，分别用 $X_1,X_2,\cdots,X_k$ 表示。令 N_i 是第 i 个类 X_i 中的变量数目，m_i 是这些变量的均值，取距离函数为欧氏距离。K-均值聚类法的步骤如下：

① 随机选择 K 个样本作为初始聚类中心 $m_1,m_2,\cdots,m_k$。

② 如果 $d(x_1,m_p)\leqslant d(x_j,m_i)$，$1\leqslant p\leqslant K$，$i=1,2,\cdots,k$，则分配 x_j 到第 p 类。

③ 重新计算每个聚类的中心：$m_i=\frac{1}{N}\sum_{x\in x_i}x$，$i=1,2,\cdots,k$。

④ 重复步骤②和③直到 m_i 不再变化，$i=1,2,\cdots,k$。

统计工具箱中实现 K-均值聚类法的函数为 kmeans。

其调用格式如下：

```
IDX = kmeans(X,k)
[IDX,C] = kmeans(X,k)
[IDX,C,sumd] = kmeans(X,k)
[IDX,C,sumd,D] = kmeans(X,k)
[...] = kmeans(...,param1,val1,param2,val2,...)
```

其中，$\boldsymbol{X}$ 是 $n\times p$ 的数据矩阵；k 是类的数目；parami，vali 等是控制迭代算法的优化参数的名称和数值；IDX 返回一个 $n\times 1$ 的向量，包含了每个变量的类编号；$\boldsymbol{C}$ 返回一个 $k\times p$ 的矩阵，表示 k 个类的中心位置；sumd 返回一个 $1\times k$ 的向量，表示每个类中所有点到聚类中心位置的距离；$\boldsymbol{D}$ 返回一个 $n\times k$ 的矩阵，表示每一个点到每一个聚类中心的距离。

【例 8-4】 将一个四维数据分成不同的类。

其实现的 MATLAB 程序代码如下：

```
>> clear all;
%产生随机数
seed=931316785;
rand('seed',seed);
randn('seed',seed);
```



```
load kmeansdata;   %装载 MATLAB 自带的数据
size(X);  %数据大小
%按照城市间的距离进行分类
%类的数目为 3
k1=3;
idx3=kmeans(X,k1,'distance','city');
%显示聚类结果
figure(1);
[silh3,h]=silhouette(X,idx3,'city');
xlabel('Silhouette 值');ylabel('聚类');
%类的数目为 4
k2=4;
idx4=kmeans(X,k2,'dist','city','display','iter');
%显示聚类结果
figure(2);
[silh4,h]=silhouette(X,idx4,'city');
xlabel('Silhouette 值');ylabel('聚类');
%类的数目为 5
k3=5;
idx5=kmeans(X,k3,'dist','city','replicates',5);
%显示聚类结果
figure(3);
[silh5,h]=silhouette(X,idx5,'city');
xlabel('Silhouette 值');ylabel('聚类');
```

运行程序，不同类数目的聚类结果分别如图 8-5~图 8-7 所示。

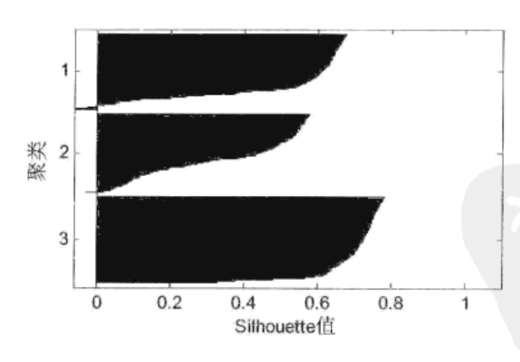


图 8-5　类数目为 3 时的聚类结果

由图 8-5 可以看出，第三类的大多数点具有较高的 silchouette 值（大于 0.6），这说明第三类与其他的类比较好地区分开了。但是第二类的许多点的 silchouette 值较低（为负值），这说明第一类和第二类没有很好地区分开。为此需要增加类的数目。

利用可选参数“display”显示算法的迭代信息如下：

```
iter  phase     num          sum
   1      1     560       2897.56
```

```
     2        1       53          2736.67
     3        1       50          2476.78
     4        1      102          1779.68
     5        1        5           1771.1
     6        2        0           1771.1
6 iterations, total sum of distances = 1771.1
```

可见最优的类数目为 4，其聚类结果如图 8-6 所示。

由图 8-6 可以看出，这 4 类很好地被分离开。继续增加类的数目为 5，得到的聚类结果如图 8-7 所示。

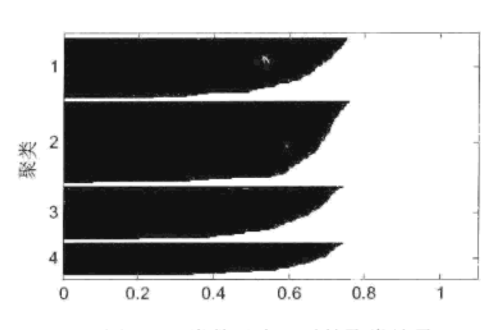


图 8-6 类数目为 4 时的聚类结果

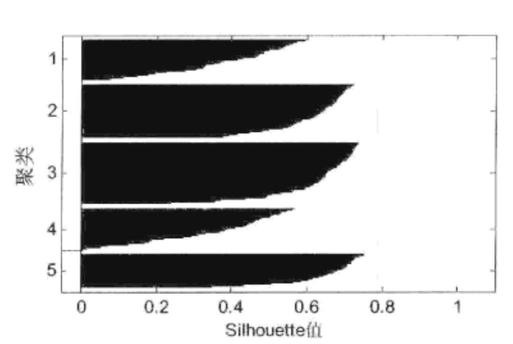


图 8-7 类数目为 5 时的聚类结果



8.3.3 聚类分析的应用示例分析

【例 8-5】 用正辛醇-水分配系数 K_{ow}、沸点 b.p.、摩尔体积 MV 和分子连接性指数 x 四个参数描述氯苯、1,4-二氯苯、五氯苯、六氯苯、4-氯硝基苯，硝基苯 6 个化合物，试根据表 8-1 中的数据对这 6 个化合物进行分类。

表 8-1 化合物性质

化合物编号	名 称	lg K_{ow}	b.p.	MV	x
1	氯苯	3.02	131.5	101.8	2.18
2	1,4-二氯苯	3.44	173.8	118.0	2.69
3	五氯苯	5.12	277.0	136.0	4.25
4	六氯苯	5.41	321.0	138.0	4.78
5	4-氯硝基苯	2.58	242.0	103.0	2.63
6	硝基苯	1.87	210.8	102.0	2.11

MATLAB 提供了两种方法进行聚类分析。

一种是一次聚类，利用函数可以对样本数据进行一次聚类，但选择面比较窄，不能更改距离的计算方法。

另一种是分布聚类，可以分以下步骤进行分布聚类：①找到数据集合中变量两两之间的相似性和非相似性，用 pdist 函数计算变量之间的距离。② 用 linkage 函数定义变量之间的连续性。③用 cophenetic 函数评价聚类信息。④用 cluster 函数创建聚类。

（1）一次聚类的 MATLAB 程序代码

```
>> x=[3.02   131.5  101.8  2.18;   3.44        173.8          118.0   2.69;   5.12          277.0            136.0
   4.25;...
   5.41      321.0   138.0   4.78;   2.58        242.0          103.0   2.63;   1.87          210.8            102.0
   2.11];
T=clusterdata(x,0.5)'
```

运行程序，输出如下：

```
T =
      2      1      3      3      4      4
```

数据集合分为 4 类。调整 cutoff 值，将有不同的分类。

（2）分布聚类的 MATLAB 程序代码

```
>> xx=zscore(x);   %数据标准化
y=pdist(xx);          %计算变量间的相似性
squareform(y);       %将输出转化为矩阵,以便阅读
z=linkage(y);         %定义变量之间的连接
c=cophenet(z,y)'     %评价聚类信息
```

运行程序，输出如下：

```
c =      0.9355
```

连接变量生成聚类树后，可以通过下列方法进行修改或了解更多的信息。

① 修改聚类树：衡量聚类信息的有效性可以用 cophenet 函数计算衡量聚类的相关性，该值越接近于 1，表示聚类效果越好。

```
>> c=cophenet(z,y)
c =      0.9355
```

将函数中距离计算方法分别指定为“Mahal”、“sEuclid”和“Cityblock”，重新计算 pdist 函数后，再用 cophenet 函数计算 c 值分别等于 0.5957、0.9355 和 0.9394，所以用“Cityblock”计算距离效果较好。

② 了解与聚类连接相关更多的信息：数据集合中聚类的方法之一是比较聚类树中每一个连接的长度与相邻次一级连接的长度。如果二者相近，则表示此水平上变量之间是相似的，这些连接被认为具有较高水平的连续性，反之，则称为不连接性的。

```
>> dendrogram(z);  %生成聚类树（见图 8-8）
```

聚类树中每一个连接的相对连续性可用 inconsistent 函数生成的不连接性系数来定量表示。该函数比较某连接的长度与相邻连接的长度的均值。若该变量与周围变量连续，则不连续性系数较低；反之，则较高。

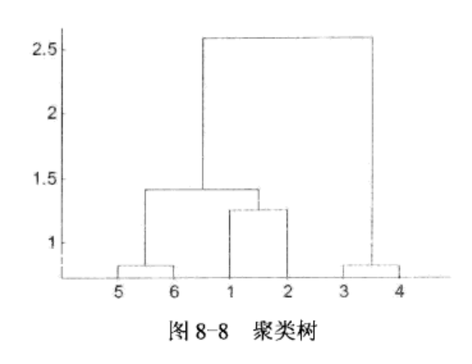


图 8-8 聚类树



```
>> I=inconsistent(z)
I =
     0.8206          0     1.0000          0
     0.8270          0     1.0000          0
     1.2539          0     1.0000          0
     1.1617     0.3056     3.0000     0.8144
     1.6076     0.8954     3.0000     1.0917
```

矩阵中，第一列为所有连接长度的均值；第二列为所有连接长度的标准偏差；第三列为计算所包含的连接数；第四列为不连续性系数。该输出信息可以与 linkage 函数的输出对照阅读。

```
>> cluster(z,0.8)'     %创建分类,以距离不超过 2 的不连续性系数为临界点
ans =
     3     3     2     2     1     1
```

从聚类树中可以清晰地了解聚类过程。比较起来，化合物 3 和 4 的性质与其他化合物相差较大。看来，苯环的氢全部或几乎全被氯取代对化合物的影响是非常显著的。

8.4 正交实验设计分析

实验设计是考虑如何安排多因素多水平的实验，能合理而高效地获得所要的分析数据，并用相应的方法分析这些数据，以确定哪些因素影响是主要的，各因素用什么水平搭配起来对实验的指标是最佳的。实验设计在改进产品分配、降低原料和能源的消耗、提高产品的产量和质量等方面具有广泛的应用。例如，缩醛化工艺是维尼纶生产的最后一道化学工艺，目的是提高维尼纶纤维的耐热水性。根据生产经验可知，反应时间、反应温度、甲醇浓度、硫酸浓度和芒硝浓度是影响产品指标的 5 个主要因素。为了寻找最佳的配方及加工工艺，芒硝浓度由于影响较小只取 3 个水平外，其他因素都取 7 个水平，如果在不同水平的组合下做全面实验，则需要 $3\times7^4=7203$ 次，而用适当的实验设计方法安排实验，可以大大减少实验次数并找到最佳配方和加工工艺，及时解决生产问题。

8.4.1 正交表分析

正交表是正交实验设计的基本工具。在正交实验设计中，安排实验，对实验结果进行分析，均在正交表上进行。下面对正交表进行较深入的介绍。

1. “完全对”与“均衡搭配”

在讲解正交表的定义和性质之前，首先介绍“完全对”与“均衡搭配”的概念。

设有两组元素 $a_1,a_2,\cdots,a_\alpha$ 与 $b_1,b_2,\cdots,b_\beta$，把 $\alpha\beta$ 个“元素对”

$$\begin{matrix} (a_1,b_1), & (a_1,b_2), & \cdots, & (a_1,b_\beta) \\ (a_2,b_1), & (a_2,b_2), & \cdots, & (a_2,b_\beta) \\ \vdots & \vdots & & \vdots \\ (a_\alpha,b_1), & (a_\alpha,b_2), & \cdots, & (a_\alpha,b_\beta) \end{matrix}$$

叫做由元素 $a_1,a_2,\cdots,a_\alpha$ 与 $b_1,b_2,\cdots,b_\beta$ 构成的“完全对”。

当不至于发生混淆时，有时也省略元素对的括号。也就是说，将 (a_i,b_j) 简写成 a_ib_j。

以后用到的“完全对”是由数码所构成的。

例如，由数码 1, 2, 3 与 1, 2, 3, 4 构成的“完全对”为

$$\begin{matrix} (1,1),(1,2),(1,3),(1,4) \\ (2,1),(2,2),(2,3),(2,4) \\ (3,1),(3,2),(3,3),(3,4) \end{matrix}$$

如果一个矩阵的某两列中，同行元素所构成的元素对（简称这两列所构成的元素对）是一个“完全对”，而且每对出现的次数相同时，称这两列“均衡搭配”；否则，称为“不均衡搭配”。

可见，所谓某两列不均衡搭配，就是指这两列所构成的元素对不是一个“完全对”；或者虽然是一个“完全对”，但并不是每个元素对出现的次数都一样。

例如，对矩阵

$$\begin{pmatrix} 1 & 1 & 2 \\ 1 & 1 & 2 \\ 1 & 2 & 1 \\ 1 & 2 & 2 \\ 2 & 1 & 2 \\ 2 & 1 & 2 \\ 2 & 2 & 2 \\ 2 & 2 & 2 \end{pmatrix}$$

其第一，二两列是均衡搭配的，因为这两列所构成的元素对是一个“完全对”，而且每对出现的次数都一样，都是两次；但是，第一，三两列为不均衡搭配，因为这两列所构成的元素对根本就不是一个“完全对”(没有元素对(2,1))；同样第二，三两列也为不均衡搭配，因为虽然这两列所构成的元素对是一个“完全对”，但并不是每个元素对出现的次数都一样，如元素对（1,1）出现一次，而元素对（1,2）却出现 3 次。显然，如果一个矩阵的第 i 列与第 j 列均衡搭配时，那么，它的第 j 列与第 i 列也必然是均衡搭配的；反之，亦然。因此，当考察了第 i,j 两列的元素对后，就不必再去考察第 j,i 两列的元素对了。

2. 正交表的定义与格式

（1）正交表的定义

有了“均衡搭配”的概念，就可以给正交表下定义了。

设 $\boldsymbol{A}$ 是一个 $n\times k$ 矩阵，它的第 j 列的元素由数码 $1,2,\cdots,t_j(j=1,2,\cdots,k)$ 所构成，如果 $\boldsymbol{A}$ 的任意两列都均衡搭配，则称 $\boldsymbol{A}$ 是一个正交表。

例如，4×3 矩阵 $\boldsymbol{A}$ 为

$$\boldsymbol{A}=\begin{pmatrix}1&1&1\\1&2&2\\2&1&2\\2&2&1\end{pmatrix}$$

该矩阵中任意两列的同行元素所构成的“元素对”都包含有 4 个数字对（1,1）、（1,2）、（2,1）、（2,2）。

这是一个“完全对”，且每个数对都出现一次，因此矩阵 $\boldsymbol{A}$ 的任何两列搭配都是均衡的，所以 $\boldsymbol{A}$ 是一张正交表。

又如：8×5 矩阵 $\boldsymbol{B}$ 为

$$\boldsymbol{B}=\begin{pmatrix}1&1&1&1&1\\1&2&2&2&2\\2&1&1&2&2\\2&2&2&1&1\\3&1&2&1&2\\3&2&1&2&1\\4&1&2&2&1\\4&2&1&1&2\end{pmatrix}$$

该矩阵第一列与其余任意列所构成的“元素对”中，都有 8 个数字对：

（1,1），（1,2），（2,1），（2,2），（3,1），（3,2），（4,1），（4,2）

这是一个“完全对”，且每个数字均出现一次；而第二、三、四、五列间的任意两列所构成的“元素对”中，都含 4 个数字对（1,1），（1,2），（2,1），（2,2）。

这是一个“完全对”，且每个数字均出现两次，所以，$\boldsymbol{B}$ 也是一张正交表。

（2）正交表的格式

在正交实验设计中，常把正交表写成表格的形式，并在其左边写上行号（实验号），在其上方写上列号（因素号）。上文提到的正交表 $\boldsymbol{A}$ 可表示为表 8-2 所示的格式，这是一张最简单的正交表。

表 8-2　正交表 $L_4(2^3)$

列号 / 行号	1	2	3
1	1	1	1
2	1	2	2
3	2	1	2
4	2	2	1

为了使用方便和便于记忆，正交表的名称一般简记为

$$L_n(m_1 \times m_2 \times \cdots \times m_k)$$

其中，L 为正交表代号（Latin 的第一个字母），n 代表正交表的行数或部分实验组合处理数，即用正交表安排实验时，应实施的实验次数。$m_1 \times m_2 \times \cdots \times m_k$ 表示正交表共有 k 列（最多可安排 k 个因素），每列水平数分别为 $m_1, m_2, \cdots, m_k$。

任何一个正交表 $L_n(m_1 \times m_2 \times \cdots \times m_k)$ 都有一个对应的具体表格。L_n 简明易记，表格则用于安排实验方案和进行实验结果分析。

3. 正交表的分类及特点

（1）等水平正交表

在正交表 $L_n(m_1 \times m_2 \times \cdots \times m_k)$ 中，若 $m_1 = m_2 = \cdots = m_k$，则称为等水平正交表，简记为 $L_n(m^k)$。式中，n 为实验次数，m 为因素的水平数，k 为正交表的列数，即最多可安排的因素数。表 8-2 所示的正交表可简记为 $L_4(2^3)$。常用的等水平正交表如下：

二水平表：$L_4(2^3)$，$L_8(2^7)$，$L_{16}(2^{15})$

三水平表：$L_9(3^4)$，$L_{27}(3^{13})$，$L_{81}(3^{40})$

四水平表：$L_{16}(4^5)$，$L_{64}(4^{21})$，…

五水平表：$L_{25}(5^6)$，$L_{125}(5^{31})$，…

等水平正交表分为标准表和非标准表两类。上面列出的都是标准表，标准表具有以下特点：

① 标准表的结构特点：

$$\begin{cases} n_i = m^{1+i}, \\ k_i = \dfrac{n_i - 1}{m-1} = \dfrac{m^{1+i} - 1}{m-1} \end{cases}, \quad i = 1, 2, \cdots$$

② 水平数相同的标准表，任意两个相邻表具有以下关系：

$$\begin{cases} n_{i+1} = mn_i, \\ k_{i+1} = n_i + k_i, \end{cases} \quad i = 1, 2, \cdots$$

显然，只要水平 m 确定了，第 i 张标准正交表就随之确定了。因此，m 是构造标准正交表的重要参数。对于任何水平的标准表，当 $i = 1$ 时，都确定了最小号正交表。

③ 利用标准表可以考察因素间的交互作用。

非标准正交表是为了缩小标准表实验号的间隔而提出来的。常用的非标准表如下：

二水平表：$L_{12}(2^{11})$，$L_{20}(2^{19})$，$L_{24}(2^{23})$，…

其他水平表：$L_{18}(3^7)$，$L_{32}(4^9)$，$L_{50}(5^{11})$，…

非标准正交表虽然为等水平表，但却不能考察因素间的交互作用。在实验中，如想考察因素间的交互作用，不能选用此类表安排实验。

（2）混合水平正交表

在正交表 $L_n(m_1 \times m_2 \times \cdots \times m_k)$ 中，如果 $m_1, m_2, \cdots, m_k$ 不完全相等，则称为混合水平正交

表。其中，最常用的是 $L_n(m_1^{k1} \times m_2^{k2})$ 混合正交表。式中，m_1^{k1} 表示水平数为 m_1 的有 $k1$ 列，m_2^{k2} 表示水平数为 m_2 的有 $k2$ 列。用这类正交表安排实验时，水平数为 m_1 的因素最多可安排 $k1$ 个，水平数为 m_2 的因素最多可安排 $k2$ 个。如前述的 8×5 矩阵 $\boldsymbol{B}$ 就是一张混合型正交表，可简记为 $L_8(4\times2^3)$。此表可安排一个四水平因素和三水平因素。



常用混合型正交表如下：

$L_8(4\times2^4)$；

$L_{12}(3\times2^4)$，$L_{12}(6\times2^2)$；

$L_{16}(4\times2^{12})$，$L_{16}(4^2\times2^9)$；

$L_{16}(4^3\times2^6)$，$L_{16}(4^4\times2^3)$，…

用混合型正交表一般不能考察交互作用，但由标准表通过并列法改造来的混合型正交表（如 $L_8(4\times2^4)$ 由 $L_8(2^7)$ 并列得到，$L_{16}(4\times2^{12})$，$L_{16}(4^2\times2^9)$ 等由 $L_{16}(2^{15})$ 并列得到），可以考察交互作用，但必须回到原标准表上进行。

4. 正交表的基本性质

由正交表的定义，可得出正交表具有下列性质。

（1）正交性

正交表正交性的主要内容是：

① 在任一列中各水平都出现，且出现的次数相等。

② 任何两列之间，各种不同水平的所有可能组合都出现，且出现的次数相等。

上述两条是判断一个正交表是否具有正交性的必要条件。

由正交表的正交性可以看出：

① 正交表的各列的地位是平等的，表中各列之间可以相互置换，称为列间置换。

② 正交表各行之间也可相互置换，称为行间置换。

③ 正交表的同一列的水平数也可以相互置换，称为水平置换。

上述 3 种置换称为正交表的 3 种初等变换。经过初等变换所得到的正交表，称为原正交表的等价表。在实际应用时，可根据不同的实验要求，把一个正交表变换成与之等价的其他特殊形式的正交表。

（2）代表性

正交表的代表性有两方面的含义。一方面，由于正交表的正交性：① 任意一列的各水平都出现，使得部分实验中包含了所有因素的所有水平。② 任意两列的所有水平都出现，使得对任意两个因素的所有水平信息及任意两因素间的所有组合信息无一遗漏。这样，虽然正交表安排的只是部分实验，但却能了解到全面实验的情况，在这个意义上，部分实验可以代表全面实验。

另一方面，由于正交表的正交性，正交实验的实验点必然均衡地分布在全面实验点中，具有很强的代表性。因此，部分实验寻找的最优条件与全面实验所找的最优条件，应有一致的趋势。

（3）综合可比性

由于正交表的正交性：① 任意一列各水平出现的次数相等。② 任意两列间所有水平组

合出现的次数相等，使得任意因素各水平的实验条件相同。这保证了在每列因素各水平的效果中，最大限度地排除了其他因素的干扰，从而可以综合比较该因素不同水平对实验指标的影响情况。这种性质称为综合可比性。

在正交表的 3 个性质中，正交性是核心，是基础；代表性和综合可比性是正交性的必然结果，从而使正交表得以具体应用。

8.4.2 不考虑交互作用正交实验设计的基本程序分析

在正交实验中有“不考虑交互作用正交实验设计”和“考虑交互作用正交实验设计”两种基本程序分析。本书只对“不考虑交互作用正交实验设计”进行介绍，关于“考虑交互作用正交实验设计”的内容，有兴趣的读者请参考相关资料。

正交实验设计的基本程序包括实验方案设计及实验结果分析两大部分。

1. 实验方案设计

下面通过一个具体的示例说明实验方案设计的内容。

【例 8-6】 啤酒酵母最适自溶条件实验。

自溶酵母提取物是一种多用途食品配料。为探讨外加中型蛋白酶方法，需做啤酒酵母的最适自溶条件实验。拟通过正交实验寻找最优工艺条件。

在安排实验时，一般应考虑如下几步。

（1）确定实验指标

实验指标是由实验目的决定的，因此实验设计之前，必然明确实验的目的，对实验所要解决的问题，应有全面、深刻的理解。通过周密考虑，确定实验指标。一项实验目的，至少需要一个实验指标，而有时在同一项实验中，由于有几个不同的实验目的，相应地，需要多个实验指标。这要根据专业知识和实验要求，具体问题具体分析，合理确定实验指标。

对本例，实验目的是寻找啤酒酵母的最适自溶条件。自溶液中蛋白质含量（%）作为实验指标，蛋白质含量越高越好。

实验指标一经确定，就应该把衡量和评价实验指标的原则、标准，测定实验指标的方法及所用仪器设备等确定下来。这本身就是一项十分细致而复杂的工作。

（2）选择实验因素

选择实验因素时，首先要根据专业知识，以往研究的结论和经验教训，尽可能全面地考虑到影响实验指标的诸因素。然后根据实验要求和尽量少选因素的一般原则，从中选定实验因素。在实际确定实验因素时，应首先选取对实验指标影响大的因素，尚未完全掌握其规律的因素和未曾被考察研究过的因素。那些对实验指标影响较小的因素，对实验指标影响规律已完全掌握的因素，应当少选或不选，但要作为可控的条件因素参加实验。实验要求考察的因素必须定为实验因素，不能遗漏，并且有时列为主要因素，进行重点考察。

在某些情况下，可以考虑多安排一些因素。例如，在初步筛选因素时，在增加因素而可以不增加实验号时，都可多选定一些实验因素。

对本例，影响蛋白质含量的因素很多，最后确定酶解温度、pH 值、加酶量为实验因素，分别以 A，B，C 表示，进行 3 因素正交实验，其他因素作为实验条件处理。

（3）选取实验因素水平，列出因素水平表

根据因素水平是作量的变化还是作质的变化，可把实验因素分为数量因素和质量因

素。例如，温度、时间、原料用量等，其水平可作量的变化，属数量因素；添加剂种类、设备型号、工艺加工方法等，其水平是由特定的质（品种、牌号等）所决定的，属质量因素。对质量因素，应选的水平常常早就定下来了，譬如使用了 3 种食品添加剂，则添加剂种类这个实验因素的水平数只能取 3。而对于数量因素或希望更多了解的实验因素，可以多取水平。



从有利于实验结果的分析考虑，水平取 3 比取 2 好。这是因为 3 水平的因素与实验指标趋势图多数为二次曲线，如图 8-9 所示，二次曲线有利于呈现实验因素水平的最佳区域。如果实验指标越高越好，由图 8-9 可得出："温度最佳条件在 100~140℃之间"的结论。而二水平因素与实验结果趋势图为线性的，如图 8-10 所示，只能得到因素水平效应的趋向，很难呈现出最佳区域，由图 8-10 只能得出"温度为 140℃时的实验指标比 100℃时的高"这一结论，最佳条件是在比 140℃更高的温度，还是在 100～140℃之间，无法判断。

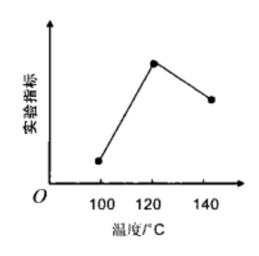


图 8-9 实验指标图 1

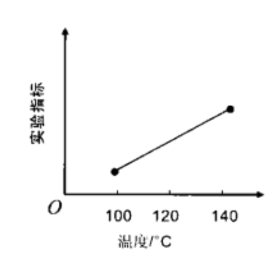


图 8-10 实验指标图 2

水平的幅度，不宜选得过宽或过窄。过窄时，实验结果可能得不到任何有用的信息；过宽，会降低实验的效率。应根据专业技术知识和已有的有关资料，尽可能把水平值取在最佳区域或接近最佳区域。如果因经验或资料不足，不能保证把水平取在最佳区域附近，则需要把水平区域拉开，尽可能使最佳区域包含在拉开的区域内。然后通过一二套实验，逐步缩小水平区，求出其最佳条件。

对例 8-6 中的各因素均选取 3 个水平，再根据专业知识和有关资料，确定每个因素的水平值。最后得到因素水平见表 8-3。

表 8-3 啤酒酵母最适自溶条件因素水平表

因素 / 水平	酵解温度/℃	pH 值	加酶量（%）
	A	*B*	*C*
1	50	6.5	2.0
2	55	7.0	2.4
3	58	7.5	2.8

（4）选择合适的正交表

确定实验因素水平后，接下来的工作就是选择一张合适的正交表。所选正交表必须满足以下条件：

① 对等水平实验，所选正交表的水平数与实验因素的水平应一致，正交表的列数应大于或等于因素及所要考察的交互作用所占的列数。

② 对不等水平实验，所选混合型正交表的某一水平的列数应大于或等于相应水平的因

素的个数。

选择正交表是一个很重要的问题。表选得太小，实验因素和要考察的交互作用就可能放不下；表选得太大，实验次数就多，不符合经济节约的原则。选正交表的原则是：在能安排下实验因素和要考察的交互作用的前提下，尽可能选择用小号正交表，以减少实验次数。另外，为考察实验误差，所选正交表安排完实验因素及要考察的交互作用后，最好有 1 空列，否则，必须进行重复实验以考察实验误差。

本例是 3 因素水平实验，可选 $L_9(3^4)$ 正交表。

（5）表头设计

正交表的每一列可以安排一个实验因素。所谓表头设计，就是将实验因素分别安排到所选正交表的各列中的过程。如果因素间无交互作用，各因素可以任意安排到正交表的各列中去；如果要考察交互作用，各因素不能任意安排，应按所选正交表的交互作用表进行安排。把因素对号入座，分别安排在正交表的各列中后，列出表头设计。对本例，表头设计见表 8-4。

表 8-4　例 8-6 表头设计

因素	A	B	C	
列号	1	2	3	4

（6）编制实验方案

在表头设计的基础上，将所选正交表中各列的水平数字换成对应因素的具体水平值，便形成了实验方案。它是实际进行实验方案的依据。

例 8-6 的实验方案见表 8-5。

表 8-5　啤酒酵母最适自溶条件实验方案及实验结果

因素	A	B	C		实验指标 Pr（%）
实验号 \ 列号	1	2	3	4	
1	1　(50)	1　(6.5)	1　(2.0)	1	6.25
2	1	2　(7.0)	2　(2.4)	2	4.97
3	1	3　(7.5)	3　(2.8)	3	4.45
4	2	1	2	3	7.53
5	2	2	3	1	5.54
6	2	3	1	2	5.50
7	3	1	3	2	11.4
8	3	2	1	3	10.9
9	3	3	2	1	8.95

表 8-5 中每个实验号对应一个组合处理，例如：

第一号实验：A_1, B_1, C_1，即酶解温度为 50℃，pH 值为 6.5，加酶量为 2.0%。

至此，实验方案设计就算完成了，随后就可以实施实验。在实验过程中，必须严格按照各号实验的组合处理进行，不能随意改动。实验因素必须严格控制，实验条件应尽量保持一致。另外，实验方案中的实验号并不意味着实际进行实验的顺序，为了加快实验，最好同时进行实验，同时取得实验结果。如果条件只允许一个一个地进行实验，为了排除外界干扰，



应使实验序列号随机化，即采用抽签、掷骰子或查随机数表的方法确定实验顺序。无论用什么顺序进行实验，一般都应进行重复实验，以减少随机误差对实验的影响。

实验结束后，将实验结果直接填入实验指标栏内，用 $x_1,x_2,\cdots,x_n$ 表示。例 8-6 的实验结果见表 8-5 的最后一栏。

正交实验的实验结果也可以不作处理，而进行"直接看"。如果由表 8-5 中的 9 个实验数据可以看出，7 号实验的蛋白质含量最高。但不能就此判断 7 号实验条件（$A_3B_1C_3$）最好。因为毕竟只做了 9 次实验，仅占 3 因素 3 水平全面实验的 1/3。不能保证全面实验中的最优组合就在所做的实验中。另一方面，还希望利用这 9 个实验数据提供的信息，了解各因素对实验指标的影响的重要程度及规律性，为此，必须对实验结果进行计算分析。

通过对实验结果的分析，可以解决以下问题：

① 分清各因素及其交互作用的主次顺序，即分清哪个是主要因素，哪个是次要因素。

② 判断因素对实验指标影响的显著程度。

③ 找出实验因素的优水平和实验范围内的最优组合，即实验因素各取什么水平时，实验指标最好。

④ 分析因素与实验指标的关系，即当因素变化时，实验指标是如何变化的。找出指标随因素变化的规律和趋势，为进一步实验指明方向。

⑤ 了解各因素之间的交互作用情况。

⑥ 估计实验误差的大小。

正交实验结果的分析方法有两种，即极差分析法（直观分析法）和方差分析法。

2. 正交实验设计的极差分析

极差分析又称直观分析法。它具有计算简便，直观形象，简单易懂等优点，是正交实验结果分析最常用的方法。

极差分析的方法简称为 R 法。它包括计算和判断两个步骤，其内容如图 8-11 所示。

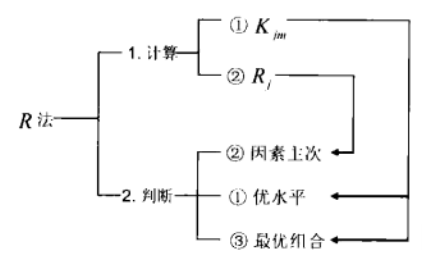


图 8-11 R 法示意图

在图 8-11 中，K_{jm} 为第 j 列因素 m 水平所对应的实验指标和。由 K_{jm} 的大小可以判断 j 列因素的优水平和各因素的优水平组合，即最优组合。

在图 8-11 中，R_j 为第 j 列因素的极差，即 j 列因素各水平下的指标最大值和最小值之差

$$R_j = \max(K_{j1}, K_{j2}, K_{jm}) - \min(K_{j1}, K_{j2}, K_{jm})$$

R_j 反映了第 j 列因素的水平变动时，实验指标的变动幅度。R_j 越大，说明该因素对实验指标的影响越大，因此也就越重要。于是依据极差 R_j 的大小，就可以判断因素的主次。

极差分析法的计算与判断可直接在实验结果分析表上进行。现以例 8-6 来说明单指标正交实验结果的极差分析法。

（1）确定因素的优水平和最优水平组合

首先分析 A 因素各水平对实验指标的影响。从表 8-5 得出，A_1 的作用只反映在 1，2，3 号实验中，A_2 的作用只反映在 4，5，6 号实验中，A_3 的作用只反映在 7，8，9 号实验中。或者说，为了考察 A_1 的作用，进行了一组实验，即由 1，2，3 号实验组成；为了考察 A_2 的作用，进行了一组实验，即由 4，5，6 号实验组成；为了考察 A_3 的作用，也进行了一组实验，即由 7，8，9 号实验组成。

A 因素 1 水平所对应的实验指标和为 $K_{A_1}=x_1+x_2+x_3=6.25+4.97+4.45=15.67$。

A 因素 2 水平所对应的实验指标和为 $K_{A_2}=x_4+x_5+x_6=7.53+5.54+5.5=18.57$。

A 因素 3 水平所对应的实验指标和为 $K_{A_3}=x_7+x_8+x_9=11.4+10.9+8.95=31.25$。

由表 8-5 可以看出，考察 A 因素进行的 3 组实验中，B,C 因素各水平都只出现了一次，且由于 B,C 间无交互作用，B,C 因素的各水平的不同组合对实验指标无影响。因此，对 A_1,A_2,A_3 来说，3 组实验的实验条件是完全一样的。如果因素 A 对实验指标无影响，那么 K_{A_1},K_{A_2},K_{A_3} 应该相等，但由上面的计算知道，K_{A_1},K_{A_2},K_{A_3} 实际上不相同，显然，这是由于 A 因素变动水平引起的，因此，K_{A_1},K_{A_2},K_{A_3} 的大小反映了 K_{A_1},K_{A_2},K_{A_3} 对实验指标影响的大小。由于蛋白质含量越大越好，而 $K_{A_1}<K_{A_2}<K_{A_3}$，所以可以判断 K_{A_3} 为 A 因素的优水平。

同理，可以计算并判断 B_1,C_1 分别为 B,C 因素的优水平。而 A,B,C 3 个因素的优水平组合 $A_3B_1C_1$ 即为本实验的最优水平组合，即加酶自溶酵母提取蛋白质含量的最优工艺条件为酶解温度为 58℃，pH 值 6.5，加酶量 2.0%。

上述 K_{jm} 的计算与优水平判断见表 8-6。

表 8-6　啤酒酵母最适自溶条件实验结果分析

因素 / 实验号	A	B	C		实验指标 Pr（%）
1	1　(50)	1　(6.5)	1　(2.0)	1	6.25
2	1	2　(7.0)	2　(2.4)	2	4.97
3	1	3　(7.5)	3　(2.8)	3	4.45
4	2　(55)	1	2	3	7.53
5	2	2	3	1	5.54
6	2	3	1	2	5.50
7	3　(58)	1	3	2	11.4
8	3	2	1	3	10.9
9	3	3	2	1	8.95
K_{j1}	15.67	25.18	22.65	20.74	
K_{j2}	18.57	21.41	21.45	21.87	
K_{j3}	31.25	18.9	21.39	22.88	$T=65.58$
优水平	A_3	B_1	C_1		
R_j	15.58	6.28	1.26		
主次顺序	ABC				



（2）确定因素主次顺序

极差 R_j 可按照上述定义计算，如 $R_A = K_{A_3} - K_{A_1} = 15.58$，同理，可计算出其他各列的极差。计算结果列于表 8-6 中。比较各 R 值可见，$R_A > R_B > R_C$，所以因素对实验指标影响的主次顺序为 ABC，即酶解温度影响最大，其次是 pH 值，而加酶量的影响最小。

（3）绘制因素水平与指标趋势图

为了更直观地反映因素对实验指标的影响规律和趋势，以因素水平为横坐标，以实验指标值（或平均值）（K_{jm}）为纵坐标，绘制因素与指标趋势图，又称关系图，如图 8-12 所示。

因素与指标趋势图可以直观地说明指标随因素水平的变化而变化的趋势，可为进一步实验时选择因素水平和指标方向。

以上即为极差分析的基本程序和方法。

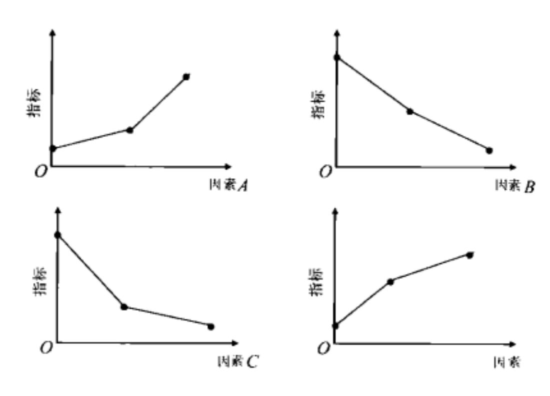


图 8-12 因素与指标趋势图

3. 正交实验设计的方差分析

正交实验设计的极差分析简单易行，计算量小，也比较直观，便于普及与推广。但是，这种方法不能把实验中由于实验条件的改变引起的数据波动同实验误差引起的数据波动区分开来。也就是说，不能区分因素各水平间对应的实验结果的差异究竟是由于因素水平不同引起的，还是由于实验误差引起的，因此不能知道实验的精度。同时，各因素对实验结果影响的重要程度，不能给予精确的数量估计，也不能提出一个标准，用来判断所考察的因素的作用是否显著。为了弥补极差分析法的不足，对正交实验结果可采用方差分析法。

（1）偏差平方和与自由度的计算

方差分析的关键是偏差平方和的分解，由前面介绍的方差分析知：

总偏差平方和与总自由度为

$$S_T = \sum_{i=1}^{n}(x_i - \overline{x})^2\text{，}\quad f_T = n-1 \tag{8-5}$$

各列偏差平方和与自由度为

$$S_j = r\sum_{i=1}^{m}(\overline{K}_{ij} - \overline{x})^2, j = 1,2,\cdots,k\text{，}\quad f_j = m-1 \tag{8-6}$$

误差偏差平方和与自由度为

$$S_e=\sum_{k_{空}}S_j\text{，}\quad f_e=\sum_{k_{空}}f_j \tag{8-7}$$

可以证明

$$S_T=\sum_{j=1}^{k}S_j=\sum_{k_{因}}S_j+\sum_{k_{交}}S_j+\sum_{k_{空}}S_j \tag{8-8}$$

$$f_T=\sum_{j=1}^{k}f_j=\sum_{k_{因}}f_j+\sum_{k_{交}}f_j+\sum_{k_{空}}f_j \tag{8-9}$$

式中，$k_{因}$，$k_{交}$，$k_{空}$ 分别为实验因素、实验考察的交互作用和空列在正交表中所占的列数，且

$$k=k_{因}+k_{交}+k_{空} \tag{8-10}$$

式（8-8）表明，总偏差平方和 S_T 等于正交表所有列的偏差平方和，等于所有实验因素、实验所考察的交互作用和空列的偏差平方和之和。式（8-9）表明，自由度 f_T 等于各列自由度之和，等于实验因素、实验所考察的交互作用和空列的自由度之和。尚需注意：①当某个交互作用占有正交表的某几列时，该交互作用的偏差平方和就等于所占各列的偏差平方和之和，其自由度也等于所占各列的自由度之和。②正交表有几个空列，误差的偏差平方和就等于所有空列的偏差平方和之和，其自由度等于所有空列的自由度之和。

现以最简单的 $L_4(2^3)$ 正交表（见表 8-7）安排的实验为例加以说明。

表 8-7　正交表 $L_4(2^3)$

列号 / 实验号	1	2	3	实验数据
1	1	1	1	x_1
2	1	2	2	x_2
3	2	1	2	x_3
4	2	2	1	x_4
K_{1j}	$K_{11}=x_1+x_2$	$K_{12}=x_1+x_3$	$K_{13}=x_1+x_4$	$T=x_1+x_2+x_3+x_4$
K_{2j}	$K_{21}=x_3+x_4$	$K_{22}=x_2+x_4$	$K_{23}=x_2+x_3$	
$\bar{K}_{1j}$	$\bar{K}_{11}=x_1+x_2/2$	$\bar{K}_{12}=x_1+x_3/2$	$\bar{K}_{13}=x_1+x_4/2$	$\bar{x}=x_1+x_2+x_3+x_4/4$
$\bar{K}_{2j}$	$\bar{K}_{21}=x_3+x_4/2$	$\bar{K}_{22}=x_2+x_4/2$	$\bar{K}_{23}=x_2+x_3/2$	

总偏差平方和

$$\begin{aligned}S_T&=\sum_{i=1}^{n}(x_i-\bar{x})^2=\sum_{i=1}^{n}x_i^2-\frac{T^2}{n}=\sum_{i=1}^{4}x_i^2-\frac{T^2}{4}\\&=x_1^2+x_2^2+x_3^2+x_4^2-\frac{1}{4}(x_1+x_2+x_3+x_4)^2\\&=\frac{3}{4}(x_1^2+x_2^2+x_3^2+x_4^2)-\frac{1}{2}(x_1x_2+x_1x_3+x_1x_4+x_2x_3+x_2x_4+x_3x_4)\end{aligned}$$

第一列各水平的偏差平方和为

$$S_1 = 2(\bar{K}_{11} - \bar{x})^2 + 2(\bar{K}_{21} - \bar{x})^2$$
$$= 2\left[\left(\frac{K_{11}}{2} - \frac{T}{2}\right)^2 + \left(\frac{K_{21}}{2} - \frac{T}{2}\right)^2\right] = \frac{1}{2}\left[(K_{11}^2 + K_{21}^2) - \frac{1}{4}T^2\right]$$
$$= \frac{1}{4}(x_1^2 + x_2^2 + x_3^2 + x_4^2) - \frac{1}{2}(x_1x_3 + x_1x_4 + x_2x_3 + x_2x_4 - x_1x_2 - x_3x_4)$$

同理，可计算出第二，三列各水平的偏差平方和S_2, S_3分别为

$$S_2 = 2(\bar{K}_{12} - \bar{x})^2 + 2(\bar{K}_{22} - \bar{x})^2 = \frac{1}{2}\left[(K_{12}^2 + K_{22}^2) - \frac{1}{4}T^2\right]$$
$$= \frac{1}{4}(x_1^2 + x_2^2 + x_3^2 + x_4^2) - \frac{1}{2}(x_1x_2 + x_1x_4 + x_2x_3 + x_3x_4 - x_1x_3 - x_2x_4)$$
$$S_3 = 2(\bar{K}_{13} - \bar{x})^2 + 2(\bar{K}_{23} - \bar{x})^2 = \frac{1}{2}\left[(K_{13}^2 + K_{23}^2) - \frac{1}{4}T^2\right]$$
$$= \frac{1}{4}(x_1^2 + x_2^2 + x_3^2 + x_4^2) - \frac{1}{2}(x_1x_2 + x_1x_3 + x_2x_4 + x_3x_4 - x_1x_4 - x_2x_3)$$

由此，可得

$$S_T = S_1 + S_2 + S_3$$

若在$L_4(2^3)$正交表的第一、第二列分别安排二水平因素A,B，则S_1,S_2分别是因素A,B的偏差平方和S_A,S_B。在不考虑交互作用的情况下，空列（第三列）的偏差平方和S_3，即为误差的偏差平方和S_e，则有

$$S_T = S_A + S_B + S_e$$

可以把上例推广到一般情况：用正交表$L_n(m^k)$（表 8-7）安排实验，实验次数为n，每个因素水平数为m，每个水平做m/n次实验，实验结果为$x_1,x_2,\cdots,x_n$，令

$$T = \sum_{i=1}^{n} x_i\text{，}\quad C_T = \frac{T^2}{n}\text{，}\quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

则总偏差平方和

$$S_T = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{T^2}{n} = Q_T - C_T \tag{8-11}$$

总偏差平方和S_T是所有数据与其总平均值的偏差平方和，它反映实验数据的总波动。

列偏差平方和

$$S_j = r\sum_{i=1}^{m}(\bar{K}_{ij} - \bar{x})^2 = \frac{1}{r}\sum_{i=1}^{n} K_{ij}^2 - \frac{T^2}{n} = Q_T - C_T\text{，}\quad j = 1,2,\cdots,k \tag{8-12}$$

特别地，当$m=2$，即二水平时，式（8-6）可表示成

$$\begin{aligned} S_j &= \frac{1}{r}(K_{1j}^2 + K_{2j}^2) - \frac{T^2}{n} = \frac{2}{n}(K_{1j}^2 + K_{2j}^2) - \frac{1}{n}(K_{1j} + K_{2j})^2 \\ &= \frac{1}{n}(K_{1j} - K_{2j})^2 = \frac{R^2}{n} \end{aligned} \tag{8-13}$$

列偏差平方和S_j是第j列中各水平对应的实验数据平均值与总平均值的偏差平方和，它

反映该列水平变动所引起的实验数据的波动。若该列安排的是因素，就称 S_j 为该因素的偏差平方和；若该列安排的交互作用，就称 S_j 为该交互作用的偏差平方和；若该列为空列，则 S_j 表示由于实验误差和未被考察的某些交互作用或某条件因素所引起的波动。在正交实验设计的方差分析中，通常把空列的偏差平方和作为实验误差的偏差平方和，虽然它属于模型误差，一般比实验误差大，但用它作为实验误差进行显著性检验，可使检验结果更可靠些。

现计算例 8-6 中各列的偏差平方和：

计算各列各水平对应数据之和 K_{1j},K_{2j},K_{3j} 及其平方 $K_{1j}^2,K_{2j}^2,K_{3j}^2$，并列于表 8-8 中。

根据式（8-12），$S_j=\frac{1}{r}\sum_{i=1}^{m}K_{ij}^2-C_T$，得

$$C_T=\frac{T^2}{n}=\frac{1}{9}\times 65.58^2=477.86$$

$$S_A=S_1=\frac{1}{3}\times(K_{11}^2+K_{21}^2+K_{31}^2)-C_T$$

$$=\frac{1}{3}(278.38+344.84+976.56)-477.86=55.4$$

同理可得，$S_B=S_2=6.49$，$S_C=S_3=0.31$，$S_e=S_4=0.83$。填入表 8-8 的最后一行中。

表 8-8　啤酒酵母最适自溶条件实验的偏差平方和计算表

表头设计	A	B	C		实验指标 Pr（%）
K_{1j}	15.76	25.18	22.65	20.74	
K_{2j}	18.57	21.41	21.45	21.87	
K_{3j}	31.25	18.99	21.48	22.97	$T=65.58$
K_{4j}	248.38	634.03	513.02	430.15	
K_{5j}	344.84	458.39	460.10	478.30	
K_{6j}	976.56	360.62	461.39	527.62	
S_j	55.4	6.49	0.31	0.83	

（2）显著性检验

偏差平方和的大小与其自由度的大小有关，不能直接比较，需经自由度平均后方可比较。将各偏差平方和除以各自相应的自由度，即得到平均偏差平方和（即方差）。

在正交实验中，各因素或交互作用的方差等于该因素或交互作用的偏差平方和除以各自相应的自由度，即

$$V_{因}=\frac{S_{因}}{f_{因}}，\ V_{交}=\frac{S_{交}}{f_{交}}，\ V_e=\sum_{k_{空}}S_j\Big/\sum_{k_{空}}f_j$$

数学上可以证明：在"假设 H_0：某因素或某交互作用不显著"成立时，统计量

$$F=\frac{V_{因}(或V_{交})}{V_e}\sim F[f_{因}(或f_{交}),f_e] \tag{8-14}$$

即统计量服从第一自由度为 $f_{因}(f_{交})$，第二自由度为 f_e 的 F 分布。因此，可把 F 作为检验统计量。对于给定的显著性水平 α，查出临界值点 F_α，若计算出的 F 值 $F_0 \geqslant F_\alpha$，则拒绝原假设 H_0，认为该因素或该交互作用对实验结果有显著影响；若 $F_0 \leqslant F_\alpha$，则接受 H_0，认为该因素或交互作用对实验结果无显著影响。



在正交实验方差分析中，还应该注意以下问题：

① 由于进行 F 检验时，要用误差偏差平方和 S_e 及自由度 f_e，而

$$S_e = \sum_{k_{空}} S_j \text{，} \quad f_e = \sum_{k_{空}} f_j$$

因此，为进行方差分析，选正交表时应留出一定空列。当无空列，又无历史资料时，应选取更大号的正交表以造成空列；或进行重复实验，以求得 S_e；或者用误差偏差平方和中的最小者作为 S_e。

② 误差的自由度一般不应小于 2，f_e 很小，F 检验灵敏度很低，有时即使因素对实验指标有影响，用 F 检验也判断不出来。

③ 为了增大 f_e，提高 F 检验的灵敏度，在进行显著性检验之前，先把各个因素和交互作用的方差 $V_{因}$ 和 $V_{交}$ 与误差方差 V_e 进行比较。如果与误差方差的大小相近，说明该因素或交互作用对实验结果的影响微乎其微，其偏差平方和是由随机误差引起的，因此可并入误差偏差平方和 S_e 中。通常把满足

$$V_{因}(\text{或}V_{交}) \leqslant 2V_e$$

的因素或交互作用的偏差平方和，并入误差偏差平方和 S_e 中，而得到新的误差偏差平方和 S_e^Δ，相应的自由度也并入 f_e 中，而得到 f_e^Δ，然后用

$$F = \frac{S_{因}(\text{或}S_{交})/f_{因}(\text{或}f_{交})}{S_e^\Delta / f_e^\Delta} \sim F[f_{因}(\text{或}f_{交}), f_e^\Delta] \tag{8-15}$$

对其他因素或交互作用进行检验。这样，使误差偏差平方和的自由度 f_e 增大，可提高 F 检验的灵敏度。

8.4.3 正交实验设计分析的应用示例分析

【例 8-7】 某化工厂生产一种产品，产率较低。现在希望通过实验设计，找出好的生产方案，以提高产率。影响产率的因素见表 8-9。

表 8-9 因素与水平

水平	因素		
	A (反应温度/℃)	B (加碱量/kg)	C (催化剂种类)
1	80	35	甲
2	85	48	乙
3	90	55	丙

解： 根据影响因素及每个因素的水平数，选择 L_9 正交表安排实验，得到的实验结果见表 8-10。对表中的数据进行分析，可得 T、$\bar{T}$ 和 R。其中，T 为各因素同一水平的结果之和，$\bar{T}$ 为其平均值，R 为极值。

表 8-10　实验结果

实验号	因素							
	A		B		C		空白列	实验结果
1	1	(80)	1	(35)	1	(甲)	1	51
2	1		2	(48)	2	(乙)	2	71
3	1		3	(55)	3	(丙)	3	58
4	2	(85)	1		2		3	82
5	2		2		3		1	69
6	2		3		1		3	59
7	3	(90)	1		2		2	77
8	3		2		1		3	85
9	3		3		2		1	84

其实现的 MATLAB 程序代码如下：

```
>> clear all;
data=[1 1 1 51;1 2 2 71;1 3 3 58;2 1 2 82;2 2 3 69;2 3 1 59;3 1 2 77;3 2 1 85;3 3 2 84];
f=3;r=3;
[r1,c]=size(data);
t=zeros(f,r);
for k=1:f
    for j=1:r
        b=0;
        for i=1:r1
            if data(i,j)==k     %水平相同
                b=b+data(i,c);
            end
        end
        t(k,j)=b;
    end
end
t1=t/3;
r=max(t1)-min(t1);
t, t1, r,                %输出结果
```

运行程序，输出如下：

```
t =
   180   269   136
   210   225   237
   246   142   263
t1 =
   60.0000   89.6667   45.3333
   70.0000   75.0000   79.0000
   82.0000   47.3333   87.6667
r =
   22.0000   42.3333   42.3333
```

从结果中可看出，理论上最优方案为$A_3B_2C_2$，最大的影响因素为A，即反应温度。

直观分析虽然比较简便易懂，但不能估计试验误差的大小，很难断定因素的重要性。为了克服这个缺点，可采用方差分析的方法。



```
>> g={[1 1 1 2 2 2 3 3 3];[1 2 3 1 2 3 1 2 3];[1 2 3 2 3 1 2 1 2]};
>> anovan(data(:,c)',g)'  %多因素方差分析
```

运行程序，输出如下（效果见图 8-13）：

```
ans =
     0.2667     0.7410     0.6427
```

Figure 1: N-Way ANOVA

File Edit View Insert Tools Desktop Window Help

Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
X1	728	2	364	2.75	0.2667
X2	92.533	2	46.267	0.35	0.741
X3	147.2	2	73.6	0.56	0.6427
Error	264.8	2	132.4		
Total	1238	8			

Constrained (Type III) sums of squares.

图 8-13 多因素效果图 1

因为 3 个因素的P值都大于 0.05，不能断定 3 个因素都不显著，而是要剔除一个最不显著的因素。在此例中剔除B，然后再作方差分析。

```
>> g1={[1 1 1 2 2 2 3 3 3];[1 2 3 2 3 1 3 1 2]};
>> anovan(data(:,4)',g1)'  %多因素方差分析
```

运行程序，输出如下（效果见图 8-14）：

```
ans =
     0.0407     0.1302
```

从而可确定A是重要因素，C是次要因素。

Figure 1: N-Way ANOVA

File Edit View Insert Tools Desktop Window Help

Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
X1	728	2	364	7.91	0.0407
X2	326	2	163	3.54	0.1302
Error	184	4	46		
Total	1238	8			

Constrained (Type III) sums of squares.

图 8-14 多因素效果图 2

【例 8-8】 在降低柴油机耗油率的研究中，根据专业人员的分析，影响因素有 4 个主要因素和水平，见表 8-11。现每个因素取两个水平做实验，并且认为因素A与B之间，A

与 C 之间可以存在交互作用。请设计实验，找出好的因素搭配，降低柴油机的耗油率。

表 8-11 因素水平表

因素	名称	单位	I 水平	II 水平
A	喷嘴器的喷嘴形式	类型	I	II
B	喷油泵柱塞直径	mm	16	14
C	供油提前角度	(°)	30	33
D	配气相位	(°)	120	140

解：在本实验中共有 4 个二水平因素，初步适用 $L_8(2^7)$ 正交表。

从 $L_8(2^7)$ 正交表的交互作用表可设计表头，安排实验并得出结果，见表 8-12。

表 8-12 实验结果

实验号	*A*	*B*	*A*×*B*	*C*	*A*×*C*	*D*	空白	实验结果
	1	2	3	4	5	6	7	*y*
1	1	1	1	1	1	1	1	228.6
2	1	1	1	2	2	2	2	225.8
3	1	2	2	1	1	2	2	230.2
4	1	2	2	2	2	1	1	218.0
5	2	1	2	1	2	1	2	220.8
6	2	1	2	2	1	2	1	215.8
7	2	2	1	1	2	2	1	228.5
8	2	2	1	2	1	1	2	214.8

其实现的 MATLAB 程序代码如下：

```
>> clear all;
x=[1 1 1 1 1 1 1 228.6;1 1 1 2 2 2 2 225.8;1 2 2 1 1 2 2 230.2;1 2 2 2 2 1 1 218.0;...
   2 1 2 1 2 1 2 220.8;2 1 2 2 1 2 1 215.8;2 2 1 1 2 2 1 228.5;2 2 1 2 1 1 2 214.8];
f=2;r=7;
[r1,c]=size(x);
t=zeros(f,r);
for k=1:f
    for j=1:r
        b=0;
        for i=1:r1
            if x(i,j)==k     %水平相同
                b=b+x(i,c);
            end
        end
        t(k,j)=b;
    end
end
T̄ =t/4,R=max(t/4)-min(t/4)
```

运行程序，输出如下：

```
T̄ ==
  225.6500   222.7500   224.4250   227.0250   222.3500   220.5500   222.7250
```



```
    219.9750   222.8750   221.2000   218.6000   223.2750   225.0750   222.9000
R =
    5.6750    0.1250    3.2250    8.4250    0.9250    4.5250    0.1750
```

从结果中可看出第八号实验 $A_2B_2C_2D_1$ 效果最好，其中因素 C 影响最大。

下面进行方差分析：

```
>> g={[1 1 1 1 2 2 2 2];[1 1 2 2 1 1 2 2];[1 2 1 2 1 2 1 2];[1 2 2 1 1 2 2 1]};
>> anovan(x(:,c),g,[1 2 3 4 5 8])'   %数据向量为方差分析的编码
```

运行程序，输出如下（效果见图 8-15）：

Figure 1: N-Way ANOVA

File Edit View Insert Tools Desktop Window Help

Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
X1	64.411	1	64.411	1051.61	0.0196
X2	0.031	1	0.031	0.51	0.6051
X1*X2	20.801	1	20.801	339.61	0.0345
X3	141.961	1	141.961	2317.73	0.0132
X1*X3	1.711	1	1.711	27.94	0.119
X4	40.951	1	40.951	668.59	0.0246
Error	0.061	1	0.061		
Total	269.929	7			

Constrained (Type III) sums of squares.

图 8-15 多因素效果图 1

```
ans =
    0.0196    0.6051    0.0345    0.0132    0.1190    0.0246
```

从方差分析可断定 B 因素最不显著，剔除 B 再作方差分析。

```
>> g1={[1 1 1 1 2 2 2 2];[1 1 2 2 2 2 1 1];[1 2 1 2 1 2 1 2];[1 2 1 2 2 1 2 1];[1 2 2 1 1 2 2 1]};
>> anovan(x(:,c),g1)'   %将 A×B 和 A×C 作为 AB、AC 看待
```

运行程序，输出如下（效果见图 8-16）：

```
ans =
    0.0007    0.0022    0.0003    0.0260    0.0011
```

Figure 1: N-Way ANOVA

File Edit View Insert Tools Desktop Window Help

Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
X1	64.411	1	64.411	1392.68	0.0007
X2	20.801	1	20.801	449.76	0.0022
X3	141.961	1	141.961	3069.43	0.0003
X4	1.711	1	1.711	37	0.026
X5	40.951	1	40.951	885.43	0.0011
Error	0.093	2	0.046		
Total	269.929	7			

Constrained (Type III) sums of squares.

图 8-16 多因素效果图 2

从而可判断出影响因素的大小顺序为 $C>A>D>A\times B>A\times C$，最好搭配为 $A_2B_1C_2D_1$。其中，B_1 由 $A\times B$ 和 $A\times C$ 的水平搭配表求出。

8.5 多元方差分析

8.5.1 多元方差分析的理论介绍

与一元统计学中的方差分析（参阅第 6 章）类似，多元样本也可以进行方差分析。两者的区别在于，一元方差分析中要分析的指标是一元随机变量，而多元方差分析中要分析的指标是多元随机变量。

8.5.2 多元方差分析的函数介绍

1）统计工具箱中实现单因素多元方差分析的函数为 manova1。

其调用格式如下：

```
d = manova1(X,group)
d = manova1(X,group,alpha)
[d,p] = manova1(...)
[d,p,stats] = manova1(...)
```

其中，X 是一个 $m\times n$ 的数值矩阵，每一行是 n 个变量的一次观测；group 是组变量，一般是一个向量或字符串数组，每一组的观测值表示来自一个总体的一个样本；alpha 是显著性水平；d 返回包含每组均值的空间维数的估计值，如果 d=0，则认为每一组的均值是同一个 n 维的多元向量；如果 d=1，则拒绝上述假设；如果 d=2，则认为多元均值位于 n 维空间内的同一个平面上，而不是在同一条直线上；p 返回均值位于 0 维、1 维空间的假设检验值，如果 p 的第 i 个分量值接近于 0，则组均值位于 $i-1$ 维空间的假设不成立。

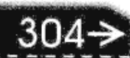
【例 8-9】 利用多元方差分析检验不同国家生产的汽车的 4 种性能参数的平均值是否存在差异。

其实现的 MATLAB 程序代码如下：

```
>> clear all;
load carbig        %装载 MATLAB 自带的数据
%多元方差分析
%分析数据矩阵
X=[MPG,Acceleration,Weight,Displacement];
%分组变量
group=Origin;
%输出结果
[d,p]=manova1(X,group)
```

运行程序，输出如下：

```
d =        3
```



```
p =      0
    0.0000
    0.0075
    0.1934
```

由于 4 种性能指标构成一组，因此组均值肯定在一个 4 维空间中。通过多元方差分析发现，实际上组均值位于 3 维子空间中，这说明 4 种性能指标的均值不尽相同。

2）统计工具箱中实现分组聚类的函数为 manovac1uter。

其调用格式如下：

```
manovacluster(stats)
manovacluster(stats,method)
H = manovacluster(stats,method)
```

其中，stats 为进行多元分析后，生成组均值的树形图；method 为指定的方法进行分类；H 为返回图中直线的句柄向量。

【例 8-10】 利用多元方差分析检验不同国家生产的汽车的 4 种性能参数的分组聚类的差异。

其实现的 MATLAB 程序代码如下：

```
>> load carbig
X = [MPG Acceleration Weight Displacement];
[d,p,stats] = manova1(X,Origin);
manovacluster(stats)
```

运行程序，效果如图 8-17 所示。

3）统计工具箱中实现均值或其他估计的多元比较检验的函数为 multcompare。

其调用格式如下：

```
c = multcompare(stats)
c = multcompare(stats ,'displayopt','ctype','estimate')
```

其中，stats 为结构中的信息进行多元比较的检验；*c* 为比较返回的结果矩阵；'displayopt'、 'ctype'、'estimate'指定进行比较的估计，并指定临界值。

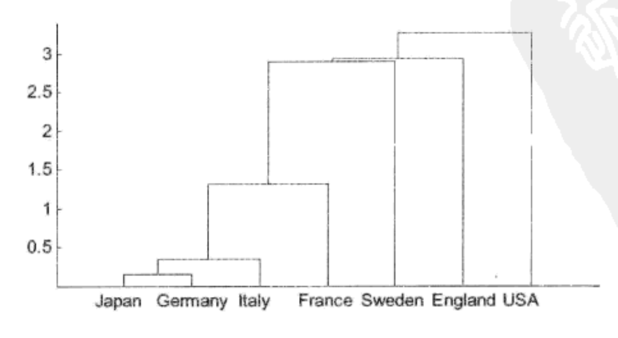


图 8-17　分组聚类效果

【例 8-11】 multcompare 函数示例。

其实现的 MATLAB 程序代码如下：

```
>> clear all;
strength = [82 86 79 83 84 85 86 87 74 82 78 75 76 77 79 79 77 78 82 79];
alloy = {'st','st','st','st','st','st','st','st',...
         'al1','al1','al1','al1','al1','al1',...
         'al2','al2','al2','al2','al2','al2'};
[p,a,s] = anova1(strength,alloy);
[c,m,h,nms] = multcompare(s);
[nms num2cell(c)]
```

运行程序，输出如下（效果见图 8-18）：

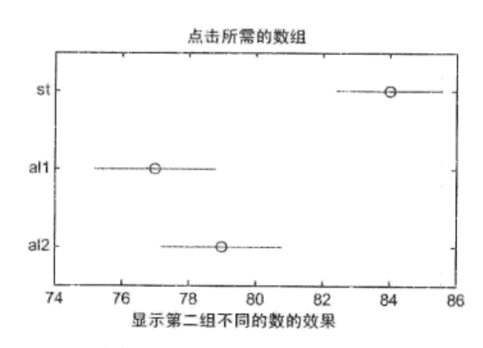


图 8-18 multcompare 效果图

```
ans =
    'st'     [1]    [2]    [ 3.6064]    [ 7]    [10.3936]
    'al1'    [1]    [3]    [ 1.6064]    [ 5]    [ 8.3936]
    'al2'    [2]    [3]    [-5.6280]    [-2]    [ 1.6280]
```

8.5.3 多元方差分析的应用示例分析

【例 8-12】 统计工具箱自带的数据文件 carsmall 是 1970 年、1976 年和 1982 年生产的不同类型汽车的性能参数测试数据。下面通过多元方差分析检验汽车的性能参数是否随时间发生了改变。

其实现的 MATLAB 程序代码如下：

```
>> clear all;
load carsmall   %装载 MATLAB 自带的数据
whos   %显示数据包含的内容
%显示变量
x=[MPG,Horsepower Displacement Weight];
gplotmatrix(x,[],Model_Year,[],'+xo');
[d,p]=manvoa1(x,Model_Year)   %多元方差分析
```

运行程序，输出如下：



原始数据包含的内容有

Name	Size	Bytes	Class	Attributes
Acceleration	100x1	800	double	
Cylinders	100x1	800	double	
Displacement	100x1	800	double	
Horsepower	100x1	800	double	
MPG	100x1	800	double	
Model	100x36	7200	char	
Model_Year	100x1	800	double	
Origin	100x7	1400	char	
Weight	100x1	800	double	

4 种性能参数变量的分组显示如图 8-19 所示。

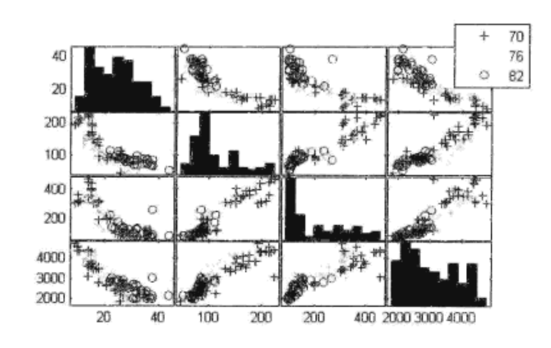


图 8-19　4 种性能参数变量的分组显示

由图 8-19 似乎可以看出，不同年份生产的汽车的性能参数有明显差别，但经过多元方差分析后的结果为：

```
d =      2
p =   1.0e-006 *
           0
     0.1141
```

即计算得到组均值的维数为 2，而不是 3。这说明其中两年生产的汽车的性能与第三年生产的汽车有明显差别。

8.6　判别分析

8.6.1 判别分析概述

1. 判别分析的基本思想及意义

在科学研究中，经常会遇到这样的问题：某研究对象以某种方式（如先前的结果或经验）划分成若干类型，而每一类型都是用一些指标 $X=(X_1,X_2,\cdots,X_p)^{\mathrm{T}}$ 来表征，即不同类型

的 X 的观测值在某种意义上有一定的差异，当得到一个新样品（或个体）的关于指标 X 的观测值时，要判断该样品（或个体）属于已知类型中的哪一个，这类问题通常称为判别分析。也就是说，判别分析是根据所研究个体的某些指标的观测值来推断该个体所属类型的一种统计方法。

判别分析的应用十分广泛。例如，在工业生产中，要根据某种产品的一些非破坏测试性测试指标判别产品的质量等级；在经济分析中，根据人均国民收入、人均农业产值、人均消费水平等指标判断一个国家的经济发展程度；在考古研究中，根据挖掘的古人头盖骨的容量、周长等判断此人的性别；在地质勘探中，根据某地的地质结构、化探和物探等各项指标来判断该地的矿化类型；在医学诊断中，医生要根据化验结果和病情征兆判断病人患了哪一种疾病，等等。值得注意的是，作为一种统计方法，判别分析所处理的问题一般都是机理不甚清楚或者基本不了解的复杂问题，如果样品的某些观测指标和其所属类型有必然的逻辑关系，也就没有必要应用判别分析方法了。

用统计的语言来描述判别分析，就是已知有 g 个总体 $G_1,G_2,\cdots,G_g$ 。每个总体 G_i 可认为是属于 G_i 的指标 $X=(X_1,X_2,\cdots,X_p)^{\mathrm{T}}$ 取值的全体，它们的分布函数 $F_1(x),F_2(x),\cdots,F_g(x)$ 均为 p 维函数，对于任一给定的新样品关于指标 X 的观测值 $x=(x_1,x_2,\cdots,x_p)^{\mathrm{T}}$ ，要判断该样品应属于 g 个总体中的哪一个。

在实际应用中，通常由取自各总体的关于指标 X 的样本作为该总体的代表，该样本称为训练样本，判别分析即取训练样本中各总体的信息以构造一定的准则来决定新样品的归属问题。训练样本往往是历史上对某现象长期观察或者使用昂贵的实验手段得到的，因此对当前的新样品，自然希望将其指标值中的信息同各总体训练样本中的信息作比较，以便在一定程度上判定新样品的所属类型。概括起来，下述几方面体现了判别分析的重要意义。

第一，为未来的决策和行动提供参考。例如，以前对一些公司在破产前两年观测到某些重要的金融指标值。现在，要根据另一个同类型公司的这些指标的观测值，预测该公司两年后是否濒临破产的危险，这便是一种判别，其结论可以帮助该公司决策人员及早采取措施，防止将来可能破产的结局。

第二，避免产品的破坏。例如，一只灯泡的寿命只有将它用坏时才能得知；一种材料的强度只有将它压坏时才能获得。一般情况下，希望根据一些非破坏性的测量指标，便可将产品分出质量等级，这也要用到判别分析。

第三，减少获得直接分类信息的昂贵代价。例如，在医学判断中，一些疾病可用代价昂贵的化验或手术得到确诊，但通常人们往往更希望通过便于观测（从而也可能导致错误）的一些外部症状来诊断，以避免过大的开支和对患者不必要的损伤。

第四，在直接分类信息不能获得的情况下可用判别分析。例如，要判断某署名的文学作品是否出自某已故作家之手，很显然，不能直接去问他。这时，可以用这位已故作家的署名作品的写作特点（用一些变量描述）作为训练样本，用判别分析方法在一定程度上判定该未署名作品是否由该作家所作。

从以上例子中也可以清楚地看出，如果不是利用直接明确的分类信息来判断某新样本的归属问题，难免会出现误判的情况。判别分析的任务是根据训练样本所提供的信息，建立在某种意义下最优（如误判概率最小，或误判损失最小等）的准则来判定一个新样品属于哪一

个总体。这里主要介绍距离判别准则。

下面，首先介绍多元正态总体的参数估计问题。

2. 多元正态分布参数的估计

在工程实际中，多元正态分布 $N(\mu,\Sigma)$ 的参数 μ 和 Σ 常常是未知的，需要通过样本来估计。



设随机向量 X 服从 p 维正态分布 $N(\mu,\Sigma)$，$X_1,X_2,\cdots,X_n$ 为来自 X 的样本 $(n>p)$，在此每个 X_i $(i=1,2,\cdots,n)$ 都是 p 维随机向量，令

$$\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i \tag{8-16}$$

$$S=\sum_{k=1}^{n}(X_k-\bar{X})(X_k-\bar{X})^{\mathrm{T}} \tag{8-17}$$

称 $\bar{X}$ 为样本均值向量，S 为样本离差阵。若令 x_i 为样品 X_i 的观察值 $(i=1,2,\cdots,n)$，则 $\bar{X}$ 与 S 的观察值分别为

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i\text{，}\quad S=\sum_{k=1}^{n}(x_k-\bar{x})(x_k-\bar{x})^{\mathrm{T}}$$

定理 8-1 若 $X_1,X_2,\cdots,X_n$ 为来自总体 X 的样本，$X\sim N_p(\mu,\Sigma)$，$\Sigma>0$，则

1）$\bar{X}$ 与 $\frac{S}{n}$ 分别是 μ 和 Σ 的最大似然估计量，即 $\hat{\mu}=\bar{X}$，$\hat{\Sigma}=\frac{S}{n}$。μ 和 Σ 的最大似然估计值分别为 $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$ 与 $\frac{S}{n}=\frac{1}{n}\sum_{k=1}^{n}(x_k-\bar{x})(x_k-\bar{x})^{\mathrm{T}}$。

2）$\bar{X}$ 与 $\frac{S}{n}$ 分别是 μ 和 Σ 的一致最小方差无偏估计，而 $\bar{x}$ 与 $\frac{S}{n-1}$ 分别是 μ 和 Σ 的最小方差无偏估计值。

定理 8-2 若 $X_1,X_2,\cdots,X_n$ 为取自 p 维正态总体 $N_p(\mu,\Sigma)$ 的样本，$\bar{X},S$ 分别由式（8-16）和式（8-17）确定，则

1）$\bar{X}$ 服从正态分布 $N_p\left(\mu,\frac{1}{n}\Sigma\right)$。

2）存在相互独立的 p 维正态变量 $Y_1,Y_2,\cdots,Y_{n-1}$，$Y_i\sim N(0,\Sigma)$，$i=1,2,\cdots,n-1$，使 S 可表示为

$$S=\sum_{i=1}^{n-1}Y_iY_i^{\mathrm{T}} \tag{8-18}$$

3）$\bar{X}$ 与 S 相互独立。

8.6.2 马氏距离

“距离”是最直观的一个概念，多元分析中的许多方法都可以用距离的观点来推导。通常是首先定义样本空间中两点之间的距离，然后定义一个点到一个总体的距离（一般定义为这个点到这个总体的均值点的距离）。如何定义样本空间中两点之间的距离

呢？在n维空间中，欧氏距离是由两点间对应坐标值之差的平方和再开方，即x,y两点间的距离平方为

$$d^2(x,y)=(x_1-y_1)^2+(x_2-y_2)^2+\cdots+(x_n-y_n)^2=(x-y)^T(x-y)$$

但在判别分析中，直接采用欧氏距离不太合适，其原因是没有考虑总体分布的分散性信息。为了克服这一不足，印度统计学家马哈拉诺必斯（Mahalanobis）于 1936 年提出了“马氏距离”。什么是“马氏距离”？与欧氏距离相比，它有什么优点呢？下面用一个简单的例子来说明这两种距离概念的差别。

设有两个正态总体G_1：$N_1(\mu_1,\sigma_1^2)$和G_2：$N_2(\mu_2,\sigma_2^2)$，今有一个样品，其值在A点，如图 8-20 所示。试问 A 点距离哪个总体近一些呢？设G_1：$N_1(5,1)$，G_2：$N_2(15,2^2)$，样品$A(9,0)$。在图 8-20 中，两条正态分布密度曲线都绘制了3σ。

从欧氏距离来看，A点与总体G_1的距离平方为$(A-5)^2$，显然小于A点与总体G_2的距离平方$(A-15)^2$，亦即A点离G_1要近一些。从概率角度来看，A在$\mu_1=5$ 右侧约$4\sigma_1$处，A在$\mu_2=15$ 的左侧约$3\sigma_2$处，根据“3σ 定律”，A点不能属于G_1，而应属于G_2。这时，若用各自的方差把“距离”标准化以后，即有

$$\frac{(A-\mu_1)^2}{\sigma_1^2}=16，\quad \frac{(A-\mu_2)^2}{\sigma_2^2}=9$$

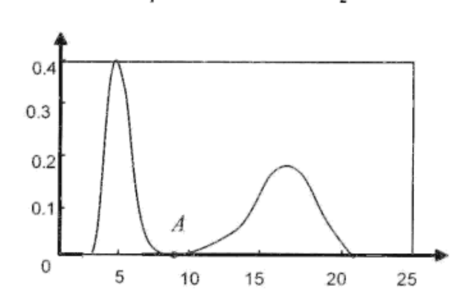


图 8-20　两条正态分布密度曲线

从而，可判断A属于G_2。推广到多维情况，就是用协方差矩阵把“距离”标准化后化为无量纲的量来作为样本空间中两点之间的距离，即定义

$$d^2(x,y)=(x-y)^{\mathrm{T}}\Sigma^{-1}(x-y)$$

这就是马氏距离。

欧氏距离还有另一个缺点，就是各个分量为不同性质的量时，“距离”的大小与单位有关。例如，点(x_1,x_2)的第一个分量x_1表示重量（以 kg 为单位），第二个分量x_2表示长度（以 cm 为单位），今有 4 个点$A(0,5)$，$B(10,0)$，$C(1,0)$，$D(0,10)$，则A与B，C与D之间的欧氏距离的平方和为

$$|AB|^2=10^2+5^2=125；\quad |CD|^2=1^2+10^2=101$$

因此AB要比CD长。

如果将点的第二个分量x_2的单位改为 mm，那么，A点的坐标就变为$(0,50)$，D点的坐标就变为$(0,100)$，B、C两点的坐标不变，这时A与B，C与D之间的欧氏距离的平

方和为

$$|AB|^2 = 10^2 + 50^2 = 2600，\ |CD|^2 = 1^2 + 100^2 = 10001$$

于是 CD 反而比 AB 长了！这显然不够合理。若用马氏距离，则与各量所用单位完全无关，就不会出现这种矛盾现象了。

下面给出同一总体下的两点间的距离，一点到一总体间的距离，以及两总体间距离的马氏定义。

定义 8-1 设 x,y 是来自总体均值向量为 μ，协方差矩阵为 Σ 的总体的两个样品，则 x,y 两点之间的马氏平方距离定义为

$$d^2(x,y) = (x-y)^{\mathrm{T}} \Sigma^{-1} (x-y) \tag{8-19}$$

定义 x 与总体 G 的马氏平方距离为

$$d^2(x,G) = (x-\mu)^{\mathrm{T}} \Sigma^{-1} (x-\mu) \tag{8-20}$$

这样，x，y 两点之间的马氏距离为

$$d(x,y) = \sqrt{(x-y)^{\mathrm{T}} \Sigma^{-1} (x-y)} \tag{8-21}$$

x 至总体 G 的马氏距离为

$$d(x,G) = \sqrt{(x-\mu)^{\mathrm{T}} \Sigma^{-1} (x-\mu)} \tag{8-22}$$

定义 8-2 设有两个总体 G_1 和 G_2，其均值向量分别是 μ_1 和 μ_2，G_1 和 G_2 的协方差矩阵相等，皆为 Σ，则总体 G_1 和 G_2 的马氏平方距离为

$$d^2(G_1,G_2) = (\mu_1-\mu_2)^{\mathrm{T}} \Sigma^{-1} (\mu_1-\mu_2) \tag{8-23}$$

可以证明，马氏距离符合通常距离的定义，即具有非负性、自反性且满足三角不等式。事实上，

$$\begin{aligned} d(x,g) &= \sqrt{d^2(x,y)} = \sqrt{(x-y)^{\mathrm{T}} \Sigma^{-1} (x-y)} \\ &= \sqrt{(x-y)^{\mathrm{T}} \Sigma^{-\frac{1}{2}} \Sigma^{\frac{1}{2}} (x-y)} \\ &= \sqrt{(\Sigma^{-\frac{1}{2}}(x-y))^{\mathrm{T}} (\Sigma^{\frac{1}{2}}(x-y))} \geqslant 0 \end{aligned}$$

仅当 $x=y$ 时，$d(x,y)=0$。

而自反性：$d(x,y)=d(y,x)$ 是很明显的。

下面求证三角不等式，设 x,y,z 为总体 G 的样品，为证明

$$d(x,z) \leqslant d(x,y) + d(y,z)$$

令

$$w = \Sigma^{-\frac{1}{2}}(x-z) = \Sigma^{-\frac{1}{2}}(x-y+y-z) = \Sigma^{-\frac{1}{2}}(x-y) + \Sigma^{-\frac{1}{2}}(y-z) \triangleq u+v$$

由 Minkowski 不等式，得

$$d(x,z) = \sqrt{w^{\mathrm{T}} w} \leqslant \sqrt{u^{\mathrm{T}} u} + \sqrt{v^{\mathrm{T}} v} = d(x,y) + d(y,z)$$

当 Σ 为单位矩阵时，马氏距离就化为通常的欧氏距离。

有了马氏距离的概念，就可以用“距离”这个尺度来判别样品的归属了。

8.6.3 多图像平均法

1）统计工具箱中实现线性判别分析的函数为 classify。

其调用格式如下：

```
class=classify(sample,traning,group)
```

其中，sample 指定数据的每一行到训练集 traning 指定的一个类中；group 指明训练集中的每一行属于哪一个类；class，它的每一个元素指定 sample 中对应元素的分类。

2）统计工具箱中实现计算马氏距离的函数为 mahal。

其调用格式如下：

```
d=mahal(Y, X)
```

其中，X 为样本至 Y 中每一个点（行）的马氏距离。

【例 8-13】 以 $\lg(1/EC_{50})$作为活性高低的界限，测定了 26 个含硫芳香族化合物对发光菌的毒性数据。分别计算了这些化合物的 $\lg K_{ow}$、Hammett 电荷效应常数σ，并测定了水解速度常数k（见表 8-13）。试根据活性类别（两类）、变量 $\lg K_{ow}$、σ 和 $\lg k$ 所取的数据，对 3 个未知活性同系物的活性进行判别。

表 8-13　26 个化合物的结构参数与判别分析结果

化合物编号与类别		$\lg(1/EC_{50})$	σ	$\lg K_{ow}$	$\lg k$
1	第 I 类(低活性)	0.93	1.28	2.30	1.76
2		10.2	0.81	3.61	2.43
3		1.03	0.81	3.81	2.31
4		1.12	1.51	3.01	1.98
5		1.13	1.04	4.32	2.20
6		1.18	1.28	0.98	1.30
7		1.32	1.28	2.30	2.05
8		1.37	1.23	0.98	1.09
9		1.41	1.04	4.32	2.12
10		1.43	1.51	1.89	1.17
11		1.45	0.81	2.29	1.48
12		1.51	1.04	3.00	1.40
13		1.51	1.48	0.95	0.57
14	第 II 类(高活性)	1.66	1.48	2.27	1.25
15		1.67	1.71	0.66	0.59
16		1.71	1.48	0.95	0.49
17		1.72	1.48	2.27	1.22
18		1.70	1.04	3.00	1.29
19		1.87	1.71	3.00	1.10
20		1.93	1.51	3.01	1.73
21		2.19	2.06	2.04	1.76
22		2.20	1.51	1.69	1.02
23		2.21	1.59	2.03	1.23
24		2.22	2.26	2.01	0.61
25		2.56	1.71	0.66	0.57
26		2.65	2.06	0.58	1.17
27	未知	1.33	0.81	2.29	1.71
28		1.72	1.59	3.35	1.46
29		1.55	1.71	3.00	1.17

其实现的 MATLAB 程序代码如下：

```
clear all;
load mydata     %保存以上数据为 mydata.mat 文件
training=[x1 x2 x3 x4]
group=[1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2]';
sample=[1.33 0.81 2.29 1.71;1.72 1.59 3.35 1.46;1.55 1.71 3.00 1.17];
class=classify(sample,training,group)'
```



运行程序，输出如下：

```
class = 1   2   2
```

即 3 个未知化合物的活性类型分别属于低、高、高，与实际结果完全一样。

8.7 实验设计分析

8.7.1 实验设计分析的理论介绍

化学实验设计和优化是数理统计方法在化学中应用比较成熟的一个领域。

实验设计是指在实验各影响因素的取值范围内，最有效地选择实验点，科学地安排实验，进而通过数据分析得到指标，取得最优值的条件的一种方法，即研究如何设计实验条件使指标获得最优值。一个良好的实验设计应能以最小的实验工作回答所有有研究对象的问题。MATLAB 中介绍了完全析因设计、不完全析因设计和 D−优化设计 3 种实验设计方法。

（1）完全析因设计

为了在几个水平上研究几个因素而设计的实验称为析因实验设计。它不仅要研究各因素水平对指标的影响，而且还强调分析诸因素对指标的作用。它按析因设计表设计方案，通过分析实验指标的变化决定各因素主效应和各因素之间的实验方法。

（2）不完全析因分析

完全析因设计的困难之一是当变量增加时，进行析因设计的组合将呈指数增长（2^n）。因此完全析因分析一般只适合于因素和水平较少的实验。当有较多因素及水平的析因分析时，可以采用不完全析因试验设计。

不完全析因分析可以通过较少的试验研究每个变量的主效应，可以大大减少实验次数。例如，当变量为 7 时，完全析因实验次数将达到 128 次，而不完全析因分析则只要 8 次。

（3）D−优化设计

不完全析因设计和熟知的正交实验，由于具有“均匀分散、整齐可比”的特点，可以用较少的实验获得各因素及其相互之间作用的丰富信息。但是，为了达到“整齐可比”的目的，往往要做较多的实验（至少为水平数的平方）。若各因素取 5 个水平，则至少要做 $5^2 = 25$ 次试验，这在实际应用中较难实现。

为此，必须寻找一种适用于多因素水平而实验次数更少的实验设计方案，20 世纪 70 年代出现的 D−优化设计便是其中的一种。D−优化设计使 Fisher 信息矩阵 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 的行列式最大化，该矩阵与参数的协方差矩阵的逆成比例，所以 det($\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$)等价于使参数的协方差矩阵的行列式最大化。

8.7.2 实验设计分析的函数介绍

统计工具箱对以上 3 种实验设计方法分别提供了相关的函数。

（1）完全析因设计相关函数

① ff2n 函数。

ff2n 函数的功能：二水平完全析因分析。

其调用格式如下：

```
X=ff2n
```

其中，X=ff2n：创建一个二水平的完全析因设计 X。

② fullfact 函数。

fullfact 函数的功能：完全析因试验设计。

其调用格式如下：

```
design=fullfact(levels)
```

其中，design=fullfact(levels)：给定因子设置，进行完全析因设计。levels 向量中的每一个元素指定 design 对应列中唯一元素的个数。

（2）不完全析因分析相关函数 fracfact

fracfact 函数的功能：生成源于生成器的不完全析因分析。

其调用格式如下：

```
x=fracfact('gen')
```

其中，x=fracfact('gen')：根据生成器字符串'gen'指定的内容生成不完全析因设计，并返回设计点的矩阵 $\boldsymbol{x}$。

（3）D-优化设计相关函数

① cordexch 函数。

cordexch 函数的功能：协同交换算法。

其调用格式如下：

```
settings=cordexch(nfactors, number)
[settings,x]=cordexch(nfactors, number)
[settings,x]=cordexch(nfactors, number,'model')
```

其中，settings=cordexch(nfactors, number)：生成因素设置矩阵 settings，number 为实验次数；[settings,x]=cordexch(nfactors, number)：生成相关的设计矩阵 $\boldsymbol{x}$；[settings,x]=cordexch(nfactors, number,'model')：为了拟合一个指定的回归模型进行设计，输入参数'model'可以是'interaction'、'quadratic'或'purequadratic'。

② daugment 函数。

daugment 函数的功能：试验设计的 D-优化扩展。

其调用格式如下：

```
settings=daugment(startdes, number)
settings=daugment(startdes, number, 'model')
```

其中，settings=daugment(startdes, number)：扩展一个初始实验设计 startdes，并进行 *n* 次新的测试；settings=daugment(startdes, number, 'model')：输入参数'model'控制回归模型的阶次。

③ dcovary 函数。

dcovary 函数的功能：用指定的协方差进行 D-优化设计。



其调用格式如下：

```
settings=dcovary(factors, covariates)
settings=dcovary(factors, covariates, 'model')
```

其中，settings=dcovary(factors, covariates)：为每一次运行创建一个有固定协变量约束的 D-优化设计；settings=dcovary(factors, covariates, 'model')：输入参数'model'控制回归模型的阶次。默认时，为一线性模型。

④ hadamard 函数。

hadamard 函数的功能：Hadamard 矩阵。

其调用格式如下：

```
H=hadamard(n)
```

其中，H=hadamard(n)：返回阶次为 *n* 的 Hadamard 矩阵。

⑤ rowexch 函数。

rowexch 函数的功能：试验设计的 D-优化设计-行交换算法。

其调用格式如下：

```
settings=rowexch(nfactors, n)
[settings, x]=rowexch(nfactors,n)
[settings, x]=rowexch(nfactors,n, 'model')
```

其中，settings=rowexch(nfactors, n)：生成因子设置矩阵 settings，用带常数项的线性累加模型进行 D-优化设计；[settings, x]=rowexch(nfactors,n)：生成相关设计矩阵；[settings, x]=rowexch(nfactors,n, 'model')：为拟合指定的回归模型生成设计。输入参数'model'控制回归模型的阶次。

8.7.3 实验设计分析的应用示例分析

【例 8-14】 请设计一个完全析因分析矩阵设置，其中有 *A*、*B* 和 *C* 3 个因素，*A* 有两个水平，而 *B*、*C* 有 3 个水平。

其实现的 MATLAB 程序代码如下：

```
>> fullfact([2 3 3])
ans = 1     1     1
      2     1     1
      1     2     1
      2     2     1
      1     3     1
      2     3     1
      1     1     2
```

```
     2     1     2
     1     2     2
     2     2     2
     1     3     2
     2     3     2
     1     1     3
     2     1     3
     1     2     3
     2     2     3
     1     3     3
     2     3     3
```

此矩阵即为实验设计设置，如第六行表示安排 A 因素的第二个水平、B 因素的第三个水平、C 因素的第一个水平作为实验点。很明显，随着因素及水平数的增多，实验次数迅速增加。

【例 8-15】 考察反应：$A \xrightarrow[\Delta]{H^+} B + C$，研究 A 的浓度及反应温度对产率的影响。请以 D-优化设计方法分析各因素的主效应和交互效应。

解：首先根据化学经验确定因素的水平，现确定 A 的浓度和反应温度各有两个水平。

考虑二输入的交互模型，使用行交换算法进行实验设计，该模型的形式为

$$y = a_0 + a_1x_1 + a_2x_2 + a_3x_1x_2 + e(\text{误差})$$

假设希望 D-优化设计通过 4 次实验来拟合模型，则

```
>> [settings,x]=rowexch(2,4,'i')
settings =
    -1     1
     1    -1
    -1    -1
     1     1
x =
     1    -1     1    -1
     1     1    -1    -1
     1    -1    -1     1
     1     1     1     1
```

据此，设计实验点，得到的实验结果见表 8-14。

表 8-14　二因素二水平 D-优化的设计表

实验序号	I	A	B	AB	指标 Y
1	+1	+1	-1	-1	80.4
2	+1	-1	-1	+1	72.4
3	+1	+1	+1	+1	94.4
4	+1	-1	+1	-1	90.6

表中第二列是为了分析各个因素对指标的影响，都以高水平表示；第三、四列为 A、B 两因素的实验点；第四列为它们之间的交互作用。

根据 D-优化设计 $\boldsymbol{X}$ 矩阵和实验结果矩阵 $\boldsymbol{Y}$，可得到系数矩阵 $\boldsymbol{A}$

$$\boldsymbol{A} = \boldsymbol{x}^{-1},\ \ \boldsymbol{Y} = n^{-1} \bullet \boldsymbol{X}^{\mathrm{T}} \bullet \boldsymbol{Y}$$



对于本题，有

```
>> x=[1 -1 -1 1;1 1 -1 -1;1 -1 1 -1;1 1 1 1];
>> y=[80.4 72.4 94.4 90.6]';
> A=1/4*x'*y
>> A'
ans =
    84.4500    -2.9500    8.0500    1.0500
```

即 A 的主效应为–2.9500，是负效应且值不大，对指标的影响小；B 的主效应为 8.0500，影响最大且为正效应；交互作用为 1.0500，影响较小，可以认为基本上不存在交互作用。当然，也可以进行方差分析以求出各影响因素。

【例 8-16】 影响分光光度法测定的因素有 pH 值、反应物、显色剂及其他掩蔽剂（或强度调节剂）浓度、反应温度等因素。请设计一个实验方案，以最少的实验次数达到最佳的结果。

解：这是一个多因素、多水平的实验，不适合采用完全析因实验设计方法。为了尽量减少实验次数，现采用 D–优化设计。

假设本例中有 m 个因素，至少做 n 次试验。先不考虑 3 次项和 3 因素间的交互作用，此时，指标函数的回归方程为

$$y = a_0 + \sum_{i=1}^{m} a_i x_i + \sum_{i=1}^{m}\sum_{j=1}^{m} x_i x_j$$

以矩阵形式表示，即为

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1T} \\ 1 & x_{21} & x_{22} & \cdots & x_{2T} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nT} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_T \end{pmatrix}$$

其中，$T = m + 0.5m(m+1)$。

应用多元线性回归分析技术，则 $\boldsymbol{B} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$,$n \geqslant T+1$

本例中 $m=4$（pH、T、反应物和显色剂浓度），所以 $T=14$，则为了满足最小二乘的条件，至少需要做 15 次左右的实验。此时，利用 rowexch 函数便可进行 D–优化设计。

```
>> [settings,x]=rowexch(4,15,'i')
```

调整实验次数，可得到不同的实验设置。根据 rowexch 函数计算结果，可得出：$11 \leqslant n \leqslant 15$。即在考虑交互作用时，至少需做 11 次实验，各实验的输入如下：

```
>> [settings,x]=rowexch(4,11,'i')
settings =
    -1    -1     1    -1
     1     1    -1    -1
     1     1    -1     1
     1     1     1    -1
     1    -1    -1     1
    -1     1    -1    -1
     1    -1     1     1
```

```
    -1     1     1     1
     1    -1     1    -1
    -1    -1    -1    -1
    -1    -1    -1     1
```

X 矩阵则是拟合上述回归方程的设计矩阵。根据 settings 矩阵安排实验点，得到指标矩阵 ***Y***，则可以根据多元回归模型（或方差分析）求出回归方程，从而找出主要影响因素和最佳实验点。

如果不考虑交互作用，则回归模型为

$$y = a_0 + \sum_{i=1}^{m} a_i x_i + \sum_{i=1}^{m} x_i^2$$

则 $T = m + m = 8$，此时只需做 9 次左右的实验，同样利用 rowexch 函数或 cordexch 函数进行优化设计：

```
>> [settings,x]=cordexch(4,9,'p')
settings =
    -1     0    -1     1
     1     1     1     1
     0     1    -1    -1
     1     0     0    -1
     0     0     1     0
     1    -1    -1     0
    -1    -1     1    -1
     0    -1     0     1
    -1     1     0     0
```

调整实验次数，比较 cordexch 函数的计算结果，可知此时最少应做 9 次实验，各次实验安排见 settings 矩阵。

【例 8-17】 试用不完全析因设计研究 5 个因子的效应。

解：对于 5 个因子的研究，如进行完全析因设计，需要 32 次实验。如果假设没有 3 因子的交互效应，则通过生成器的智能选择，发现只要经过 8 次实验就可以估计这 5 个效应。

其实现的 MATLAB 程序代码如下：

```
>> [x,c]=fracfact('a b c a b a c');
>> c(1:7,:)
ans =
    'Term'    'Generator'    'Confounding'
    'X1'      'a'            'X1 + X4 + X6'
    'X2'      'b'            'X2 + X5'
    'X3'      'c'            'X3 + X7'
    'X4'      'a'            'X1 + X4 + X6'
    'X5'      'b'            'X2 + X5'
    'X6'      'a'            'X1 + X4 + X6'
```

这里所有的主效应由一个或多个二因子交互组成，8 个实验的输入由 ***X*** 矩阵设计。



第 9 章 隐马尔可夫模型及统计工具箱的示范程序

9.1 隐马尔可夫模型

9.1.1 基本理论概述

隐马尔可夫模型（Hidden Markov Model，HMM）是马尔可夫链的一种。它的状态不能直接观察到，但能通过观测向量序列观察到每个观测向量都是通过某些概率密度分布表现为各种状态，每一个观测向量由一个具有响应概率密度分布的状态序列产生。所以，隐马尔可夫模型是一个双重随机过程——具有一定状态数的隐马尔可夫链和显示随机函数集。自 20 世纪 80 年代以来，HMM 被应用于语音识别，取得重大成功。到了 20 世纪 90 年代，HMM 还被引入计算机文字识别和移动通信核心技术“多用户的检测”。近年来，HMM 在生物信息科学、故障诊断等领域也开始得到应用。

1. 基本概念

假设某个系统由 5 个不同的状态 $(S_1, S_2, S_3, S_4, S_5)$ 组成，如图 9-1 所示。

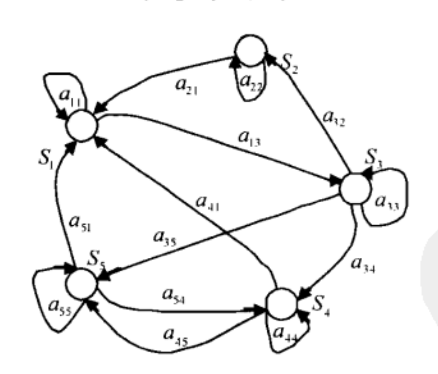


图 9-1 5 个不同状态的马尔可夫链

显然这是一个离散的马尔可夫过程，记状态变化的时间常数为 $t = 1, 2, \cdots, t$ 时刻的实际状态为 q_1。为了描述上述系统的全部概率，一般来说，需要知道当前时刻和以前时刻的状态。特别地，对一阶马尔可夫链来说，其概率描述可以简化为

$$P\{q_1 = S_j \mid q_{t-1} = S_i, q_{t-2} = S_k, \cdots\} = P\{q_1 = S_j \mid q_{t-1} = S_i\} \tag{9-1}$$

由式（9-1）等号的右边可知，概率与时间无关，因此可以写成状态转移概率的形式

$$a_{ij}=P\{q_t=S_j|_{q_{t-1}}=S_i\}，\ 1\leqslant i，\ j\leqslant N \tag{9-2}$$

且满足以下属性：

$$a_{ij}\geqslant 0$$
$$\sum_{j=1}^{N}a_{ij}=1 \tag{9-3}$$

因为上述过程的输出等于每个时刻的状态集，因此也称作是一个可观测的隐马尔可夫模型。这种模型在实际中受到很多限制，可以对其进行扩展，假定观测量是状态的概率函数，即嵌入一个不可观测的随机过程，得到的模型是一个双重随机过程，这就是隐马尔可夫模型。

一个隐马尔可夫模型具有以下元素：

- N——模型的隐状态数目。虽然这些状态是隐含的，但是在许多实际应用中，模型的状态通常有具体的物理意义。
- M——每个状态的不同观测值的数目。
- A——状态转移概率矩阵 $A=\{a_{ij}\}$，且有

$$a_{ij}=P\{q_{t+1}=S_j\mid q_t=S_i\}，\ 1\leqslant i，\ j\leqslant N$$

- B——观测概率矩阵 $B=\{b_j(k)\}$，且有

$$b_j(k)=P\{v_k\mid q_t=S_j\}，\ 1\leqslant j\leqslant N, 1\leqslant k\leqslant M$$

也就是说，当状态为 S_j 时，观测结果为 v_k 的后验概率。

- π——初始状态概率矩阵 $\pi=\{\pi_i\}$，且有

$$\pi_i=P\{q_1=S_i\}，\ 1\leqslant i\leqslant N$$

一般地，可以用 $\lambda=(A,B,\pi)$ 简洁地表示一个隐马尔可夫模型。给定了 N,M,A,B,π 后，隐马尔可夫模型可以产生一个观测序列

$$O=O_1O_2\cdots O_T$$

其过程如下：

① 根据初始状态概率矩阵 π，选择一个初状态 $q_1=S_i$。

② 令 $t=1$。

③ 根据状态 S_i 中的符号的概率密度函数 $f_i(x_t)$，选 $x_t=v_k$。

④ 根据状态 S_i 的状态转移概率 a_{ij}，转移到一个新的状态 $q_{t+1}=S_j$。

⑤ 令 $t=t+1$，若 $t<T$ 返回③，否则结束。

实际中，应用隐马尔可夫模型时必须解决 3 个基本问题：

1）给定观测序列 $O=O_1O_2\cdots O_T$ 和模型参数 $\lambda=(A,B,\pi)$，怎样有效计算观测序列的概率，即 $P\{O\mid\lambda\}$？

这实际上是一个计算问题，即给定模型参数和观测序列，如何计算由该模型产生的观测序列的概率。也可以把这个问题看做评估给定模型与给定序列的匹配程度。

2）给定观测序列 $O = O_1O_2\cdots O_T$ 和模型参数 $\lambda = (A,B,\pi)$，怎样选择一个在某种意义上最优的状态序列 $Q = q_1q_2\cdots q_T$？

这个问题试图揭示模型隐含的内容，即找到“正确”的状态序列，但实际上这是不可能的。因此，在实际问题中，经常需要利用最优准则解决这个问题。

3）怎样调整模型参数 $\lambda = (A,B,\pi)$，使其最大？

这个问题实际上是模型参数的优化问题，使其更准确地解释给定观测序列是怎样产生的。用来调整模型参数的观测序列称为训练序列，用它可以训练隐马尔可夫模型。

根据是否对观测向量进行矢量化，可把隐马尔可夫模型分为离散概率密度型和连续概率密度型。对于离散概率密度型，观测向量只能取码本中的有限个码字；对于连续概率密度型，其形式有高斯型、高斯混合型、高斯自回归型等。另外，根据不同状态之间相互转换的规律不同，可将隐马尔可夫模型分为自左到右和全连接两种拓扑结构。

2. 基本算法

针对前面 3 个基本问题，人们提出了相应的算法。

（1）前向-后向算法

这个算法用来计算给定一个观测序列 $O = O_1O_2\cdots O_T$ 和模型参数 $\lambda = (A,B,\pi)$ 时，模型的输出概率 $P\{O|\lambda\}$。

隐马尔可夫模型的前向概率为

$$a_t(i) = P\{O_1O_2\cdots O_t, q_t = i|\lambda\} \quad (9-4)$$

表示给定 HMM 参数，部分观测序列 $\{O_1O_2\cdots O_t\}$ 在时刻 t 处于状态 i 的概率。

它的递推计算公式如下：

① 初始化：

$$a_t(i) = \pi_i b_i(O_1)，\ 1 \leqslant i \leqslant N \quad (9-5)$$

② 迭代计算：

$$a_{t+1}(j) = \left(\sum_{i=1}^{N} p_t(i)a_{ij}\right)b_j(O_{t+1})，\ 1 \leqslant t \leqslant T-1，\ 1 \leqslant j \leqslant N \quad (9-6)$$

③ 终止计算：

$$P(O|\lambda) = \sum_{i=1}^{N} a_t(i) \quad (9-7)$$

隐马尔可夫模型的后向概率为

$$\beta_t(i) = P\{O_{t+1}O_{t+2}\cdots O_T, q_T = i|\lambda\} \quad (9-8)$$

表示给定 HMM 参数，观测序列在时刻 t 处于状态 i，系统输出部分观测序列 $(O_{t+1}O_{t+2}\cdots O_T)$ 的概率。

它的递推计算公式如下：

① 初始化：

$$\beta_T(i) = 1，\ 1 \leqslant i \leqslant N \quad (9-9)$$

② 迭代计算：

$$\beta_t(i)=\sum_{j=1}^{N}a_{ij}b_j(O_{t+1})\beta_{t+1}(j)\text{，}\ 1\leqslant t\leqslant T-1\text{，}\ 1\leqslant i\leqslant N \tag{9-10}$$

这样，可以根据前向概率和后向概率得到 HMM 整个观测序列的输出概率为

$$P(O|\lambda)=\sum_{i=1}^{N}a_t(i)\beta_t(i)b_j(O_{t+1})=\sum_{i=1}^{N}\sum_{j=1}^{N}a_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)\text{，}\ 1\leqslant t\leqslant T-1 \tag{9-11}$$

（2）Viterbi 算法

这个算法用于解决给定一个观测序列 $O=O_1O_2\cdots O_T$ 和模型参数 $\lambda=(A,B,\pi)$ 时，在最优的意义上确定一个状态序列 $Q=q_1q_2\cdots q_T$ 的问题。Viterbi 算法广泛应用于通信领域的动态规划，它不仅可以找到一条“最优”的状态转移路径，还可以得到该路径对应的输出概率。

Viterbi 算法可以叙述为：定义 $\delta_t(i)$ 为时刻 t 时沿一条路径，$q_1q_2\cdots q_t,q_t=S_i$，则它产生出 $O_1O_2\cdots O_t$ 的概率最大。

其过程如下：

① 初始化：

$$\begin{aligned}&\delta_1(i)=\pi_i b_i(O_1),\ 1\leqslant i\leqslant N\\&\psi_1(i)=1,\ 1\leqslant i\leqslant N\end{aligned} \tag{9-12}$$

② 递归计算：

$$\begin{aligned}&\delta_t(j)=\max_{1\leqslant i\leqslant N}[\delta_{t-1}(i)a_{ij}]b_j(O_t),\ 2\leqslant t\leqslant T,\ 1\leqslant j\leqslant N\\&\psi_t(j)=\arg\max_{1\leqslant i\leqslant N}[\delta_{t-1}(i)a_{ij}],\ 1\leqslant t\leqslant T,\ 1\leqslant j\leqslant N\end{aligned} \tag{9-13}$$

③ 终止计算：

$$\begin{aligned}&P^*=\max_{1\leqslant i\leqslant N}[\delta_T(i)]\\&q_T^*=\arg\max_{1\leqslant i\leqslant N}[\delta_T(i)]\end{aligned} \tag{9-14}$$

④ 求取状态序列：

$$q_t^*=\psi_{t+1}(q_{t+1}^*),\ 1\leqslant t\leqslant T-1 \tag{9-15}$$

（3）Baum-Welch 算法

这个算法用于解决 HMM 的参数估计问题，即给定一个观测序列 $O=O_1O_2\cdots O_T$，该算法能确定一个 $\lambda=(A,B,\pi)$，使得 $P\{O|\lambda\}$ 最大。

由式（9-11）可知，需要求取 λ，使得概率 $P\{O|\lambda\}$ 最大。实际中，Baum-Welch 算法利用递归的思想，使得 $P\{O|\lambda\}$ 局部最大，最后得到模型参数 $\lambda=(A,B,\pi)$。

记 $\xi_i(i,j)$ 为给定训练序列 O 和模型参数 λ 时，t 时刻马尔可夫链处于 S_i 状态和 $t+1$ 时刻为 S_j 状态的概率，即

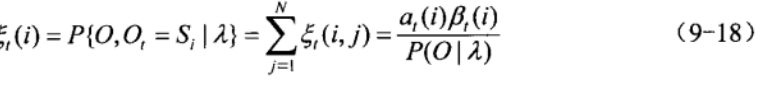
$$\xi_t(i,j)=P\{O,O_t=S_i,O_{t+1}=S_j|\lambda\} \tag{9-16}$$

可以推导出

$$\xi_t(i,j)=[a_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)]/P\{O|\lambda\} \tag{9-17}$$

于是，t 时刻马尔可夫链处于 S_i 状态的概率为





$$\xi_t(i)=P\{O,O_t=S_i|\lambda\}=\sum_{j=1}^{N}\xi_t(i,j)=\frac{a_t(i)\beta_t(i)}{P(O|\lambda)} \tag{9-18}$$

因此，$\sum_{t=1}^{T-1}\xi_t(i)$ 表示从 S_i 状态转移出去的次数的期望值，而 $\sum_{t=1}^{T-1}\xi_t(i,j)$ 表示从 S_i 状态转移到 S_j 状态的次数的期望值。由此，可以得到 Baum-Welch 算法的重估公式：

$$\overline{\pi}_i=\xi_1(i) \tag{9-19}$$

$$\overline{a}_{ij}=\frac{\sum_{t=1}^{T-1}\xi_t(i,j)}{\sum_{t=1}^{T-1}\xi_t(i)} \tag{9-20}$$

$$\overline{b}_j(O_t=v_k)=\frac{\sum_{t=1,O_t=v_k}^{T}\xi_t(j)}{\sum_{t=1}^{T}\xi_t(j)} \tag{9-21}$$

HMM 的参数估计过程如下：

① 选取初始模型参数 $\lambda=(A,B,\pi)$。

② 根据观测序列 O，由式（9-19）~式（9-21）求得一组新参数 $\overline{\lambda}=(\overline{A},\overline{B},\pi)$，可以证明 $P\{O|\overline{\lambda}\}>P\{O|\lambda\}$。

③ 重复第②步，逐步改进模型参数，直至 $P\{O|\overline{\lambda}\}$ 收敛，此时的 $\overline{\lambda}$ 即为所求的模型参数。

9.1.2 相关函数介绍

给定一个隐马尔可夫模型，要想在实际中能够应用，必然解决以上 3 个基本问题。统计工具箱提供了 5 个函数用于隐马尔可夫模型分析：

1）hmmgenerate——产生一个隐马尔可夫模型序列。

其调用格式如下：

```
[seq,states] = hmmgenerate(len,TRANS,EMIS)
hmmgenerate(...,'Symbols',SYMBOLS)
hmmgenerate(...,'Statenames',STATENAMES)
```

其中，*len* 是序列的长度；***TRANS*** 是状态转移概率矩阵；***EMIS*** 是观测概率矩阵；seq 返回一个观测序列；states 返回一个状态序列。

【例 9-1】 根据给定的状态转移概率矩阵和观测概率矩阵生成隐马尔可夫模型序列。

其实现的 MATLAB 程序代码如下：

```
>> clear all;
%状态转移概率矩阵
trans = [0.95,0.05;0.10,0.90];
%观测概率矩阵
  emis = [1/6 1/6 1/6 1/6 1/6 1/6;...
      1/10 1/10 1/10 1/10 1/10 1/2];
%产生隐马尔可夫模型序列
len=10;
[seq,states] = hmmgenerate(len,trans,emis)
[seq,states] = hmmgenerate(len,trans,emis,...
      'Symbols',{'one','two','three','four','five','six'},...
      'Statenames',{'fair';'loaded'})
```

运行程序，输出如下：

观测序列为

```
seq =
     1     6     6     3     5     1     3     6     6     6
```

状态序列为

```
states =
     1     1     1     1     1     1     1     1     2     2
```

观测序列为

```
seq =
    'five'    'one'    'two'    'one'    'one'    'five'    'five'    'two'    'six'    'one'
```

状态序列为

```
states =
    'fair'    'fair'    'fair'    'fair'    'fair'    'fair'    'fair'    'fair'    'fair'    'fair'
```

2）hmmdecode：计算给定观测序列的概率[求解第 1)个问题]。

其调用格式如下：

```
PSTATES = hmmdecode(seq,TRANS,EMIS)
[PSTATES,logpseq] = hmmdecode(...)
[PSTATES,logpseq,FORWARD,BACKWARD,S] = hmmdecode(...)
hmmdecode(...,'Symbols',SYMBOLS)
```

其中，seq 是观测序列；***TRANS*** 是状态转移概率矩阵；***EMIS*** 是观测概率矩阵；*PSTATES* 返回后验概率；*logpseq* 返回观测序列概率的对数值；*FORWARD*，*BACKWARD* 分别返回序列的前向概率和后向概率概率；*S* 是尺度。

【例 9-2】 计算给定模型下的观测序列的后验概率。

其实现的 MATLAB 程序代码如下：

```
>> clear all;
%状态转移概率矩阵
```

```
trans = [0.95,0.05;0.10,0.90];
%观测概率矩阵
emis = [1/6 1/6 1/6 1/6 1/6 1/6;
    1/10 1/10 1/10 1/10 1/10 1/2];
 %产生隐马尔可夫模型序列
 len=10;
[seq1,states] = hmmgenerate(len,trans,emis);
seq1
%计算观测序列的后验概率
[pStates1,logp1]=hmmdecode(seq1,trans,emis)
%产生隐马尔可夫模型序列
[seq2,states] = hmmgenerate(len,trans,emis,...
    'Symbols',{'one','two','three','four','five','six'});
seq2
%计算序列的后验概率
[pStates2,logp2]= hmmdecode(seq2,trans,emis,...
    'Symbols',{'one','two','three','four','five','six'})
```

运行程序，输出如下：

第一次产生的观测序列为

```
seq1 =
     3     4     5     3     4     3     6     5     2     4
```

其后验概率为

```
pStates1 =
 0.9912   0.9846   0.9781   0.9692   0.9548   0.9299   0.8857   0.9053  了 0.9079   0.8945
 0.0088   0.0154 0.0219   0.0308   0.0452   0.0701   0.1143   0.0947   0.0921   0.1055
```

观测序列概率的对数值

```
logp1 =   -18.2070
```

第二次产生的观测序列为

```
seq2 =
    'five'   'three'   'five'   'five'   'three'   'one'   'two'   'three'   'two'   'two'
```

其后验概率为

```
pStates2 =
0.9924   0.9880   0.9853   0.9833   0.9814   0.9789   0.9748   0.9678   0.9554   0.9331
0.0076   0.0120   0.0147   0.0167   0.0186   0.0211   0.0252   0.0322   0.0446   0.0669
```

观测序列概率的对数值

```
logp2 =   -18.3029
```

3）hmmestimate——给定观测序列和状态序列下，估计隐马尔可夫模型的参数。

其调用格式如下：

```
[TRANS,EMIS] = hmmestimate(seq,states)
hmmestimate(...,'Symbols',SYMBOLS)
hmmestimate(...,'Statenames',STATENAMES)
hmmestimate(...,'Pseudoemissions',PSEUDOE)
hmmestimate(...,'Pseudotransitions',PSEUDOTR)
```

其中，seq 是观测序列；states 是状态序列；TRANS 返回状态转移概率矩阵的极大似然估计；EMIS 返回观测概率矩阵的极大似然估计。

【例 9-3】 给定观测序列和状态序列下，估计隐马尔可夫模型的参数。

其实现的 MATLAB 程序代码如下：

```
>> clear all;
%状态转移概率矩阵
trans = [0.95,0.05; 0.10,0.90];
%观测概率矩阵
emis = [1/6 1/6 1/6 1/6 1/6 1/6;...
   1/10 1/10 1/10 1/10 1/10 1/2]
%产生隐马尔可夫模型序列
len=10;
[seq,states] = hmmgenerate(1000,trans,emis);
%估计模型的参数
[estimateTR,estimateE] = hmmestimate(seq,states)
```

运行程序，输出如下：

实际的状态转移概率矩阵为

```
trans =
     0.9500    0.0500
     0.1000    0.9000
```

实际的观测概率矩阵为

```
emis =
     0.1667    0.1667    0.1667    0.1667    0.1667    0.1667
     0.1000    0.1000    0.1000    0.1000    0.1000    0.5000
```

估计的状态转移概率矩阵为

```
estimateTR =
     0.9520    0.0480
     0.0961    0.9039
```

估计的观测概率矩阵为

```
estimateE =
     0.1499    0.1679    0.1589    0.1724    0.1724    0.1784
     0.0871    0.1021    0.1111    0.0931    0.1081    0.4985
```

比较可见，估计值和实际值是一致的。

4）hmmviterbi——计算隐马尔可夫模型序列的最可能的状态路径[求解第 2)个问题]。

其调用格式如下：



```
STATES = hmmviterbi(seq,TRANS,EMIS)
hmmviterbi(...,'Symbols',SYMBOLS)
hmmviterbi(...,'Statenames',STATENAMES)
```

其中，seq 是观测序列；TRANS 是状态转移概率矩阵；EMIS 是观测概率矩阵；STATES 返回最可能的状态序列。

【例 9-4】 给定观测序列和模型下，计算最可能的状态序列。

其实现的 MATLAB 程序代码如下：

```
>> clear all;
%状态转移概率矩阵
trans = [0.95,0.05;0.10,0.90];
%观测概率矩阵
emis = [1/6 1/6 1/6 1/6 1/6 1/6;...
    1/10 1/10 1/10 1/10 1/10 1/2];
%产生隐马尔可夫模型序列
len=10;
[seq,states] = hmmgenerate(len,trans,emis);
%计算状态转换路径
estimatedStates1 = hmmviterbi(seq,trans,emis)
%产生隐马尔可夫模型序列
[seq,states] =hmmgenerate(len,trans,emis,...
                    'Statenames',{'fair';'loaded'});
%计算状态转移路径
estimatesStates2 = hmmviterbi(seq,trans,emis,...
                    'Statenames',{'fair';'loaded'})
```

运行程序，输出如下：

第一次估计的最可能状态序列为

```
estimatedStates1 =
     1     1     1     1     1     1     1     1     1     1
```

第二次估计的最可能状态序列为

```
estimatesStates2 =
   'fair'  'fair'  'fair'   'fair'   'loaded'   'loaded'  'loaded'  'loaded'  'loaded'    'loaded'
```

5）hmmtrain——隐马尔可夫模型参数的极大似然估计[求解第 3）个问题]。其调用格式如下：

```
[ESTTR,ESTEMIT] = hmmtrain(seq,TRGUESS,EMITGUESS)
hmmtrain(...,'Algorithm',algorithm)
```

```
hmmtrain(...,'Symbols',SYMBOLS)
hmmtrain(...,'Tolerance',tol)
hmmtrain(...,'Maxiterations',maxiter)
hmmtrain(...,'Verbose',true)
hmmtrain(...,'Pseudoemissions',PSEUDOE)
hmmtrain(...,'Pseudotransitions',PSEUDOTR)
```

其中，seq 是观测序列；TRGUESS 是状态转移概率矩阵的初始值；EMITGUESS 是观测概率矩阵的初始值；ESTTR 返回状态转移概率矩阵的极大似然估计；ESTEMIT 返回观测概率矩阵的极大似然估计。

【例 9-5】 利用给定的观测序列对模型进行训练。

其实现的 MATLAB 程序代码如下：

```
>> clear all;
%状态转移概率矩阵
trans = [0.95,0.05;0.10,0.90];
%观测概率矩阵
emis = [1/6 1/6 1/6 1/6 1/6 1/6;...
    1/10 1/10 1/10 1/10 1/10 1/2];
%产生隐马尔可夫模型序列
len=100;
seq1 = hmmgenerate(len,trans,emis);
trans,emis
seq2 = hmmgenerate(2*len,trans,emis);
seqs = {seq1,seq2};
[estTR,estE] = hmmtrain(seqs,trans,emis)
```

运行程序，输出如下：

初始状态转移概率矩阵为

```
trans =
    0.9500    0.0500
    0.1000    0.9000
```

初始观测概率矩阵为

```
emis =
    0.1667    0.1667    0.1667    0.1667    0.1667    0.1667
    0.1000    0.1000    0.1000    0.1000    0.1000    0.5000
```

训练后的状态转移概率矩阵为

```
estTR =
    0.9854    0.0146
    0.0342    0.9658
```

训练后的观测概率矩阵为

```
estE =
```

0.1424	0.1824	0.1797	0.1390	0.1731	0.1835
0.1339	0.0563	0.1220	0.0956	0.1268	0.4654



9.1.3 HMM在语音识别中的应用

1. 基本原理

隐马尔可夫模型（HMM）在语音识别中的作用非常重要，得到了广泛的应用。在语音识别中，首先需要建立一种对应关系，比如，使一个字对应一个HMM，此时模型的状态就对应这个字所包含的全部可能的音素。对应于该字的一个观测样本，这些音素会按照一定的顺序出现，这样就形成了HMM中的状态序列，是实际中不可观测的。实际中，可以观测每个字母声信号的振幅。为了建立上述对应关系，首先需要对该字的一组观测样本进行学习，也就是进行HMM参数估计。

学习了每个字的参数后，就可以用于识别。也就是对任意的一组观测样本，找到最大可能产生该观测样本的模型作为该字的代表。一个典型的语音识别过程如图9-2所示。

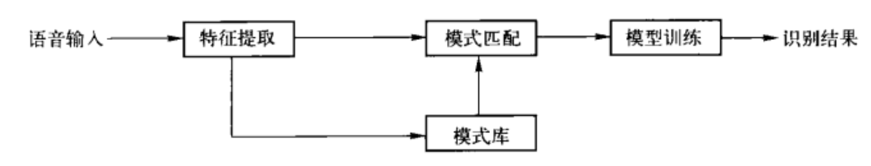


图9-2 语音识别的基本过程

由此可见，语音识别过程主要包括特征提取、模式匹配及模型训练3个方面。此外，还涉及语音识别单元的选取。

（1）语音识别单元的选取

选择识别单元是语音识别应用的第一步。语音识别单元包括单词、音节和音素3种，实际中根据具体的应用来选取。

（2）特征提取

语音信号中含有丰富的信息，但如何从中提取出对语音识别有用的信息是一个关键的问题。通过特征提取，可以对语音信号进行分析处理、去除无关紧要的冗余信息，获得语音信号的重要信息。

（3）模式匹配及模型训练

HMM是语音信号识别特征的参数表示，它由相关联的两个随机过程共同描述信号的统计特性。其中，一个是隐含的具有有限个状态的马尔可夫链（不可观测）；另一个是与马尔可夫链的每一个状态相关联的观察矢量的随机过程（可观测）。其中，隐马尔可夫链的特征要依赖观测到的信号特征来揭示。于是，语音信号的某一段的特征就由对应状态的观察符号的随机过程描述，而信号随时间的变化由隐马尔可夫链的状态转移概率矩阵描述。

模型训练是按照一定的准则，从大量已知模式中获取表征该模式特征的模型参数。模式匹配则是根据一定的准则，使未知模式与模型库中的某一个模型获得最佳匹配。

2. 示例分析

在此，将通过一个具体的例子来说明独立词的语音识别步骤，具体背景为：利用HMM识别单词0~9，每个单词都有重复的10次发音，每一个发音的语音信号的长度为4800。

（1）信号预处理

将采集的语音信号分成长度为 N 的块，相邻块起点之间的间隔为 ΔN 。比如，长度为 $N_s = 10000$ 的样本，取 $N = 320$ ， $\Delta N = 80$ ，则块的数目为 $T = 1 + [(N_s - N)/\Delta N] = 122$ 。这样，观测时间可以表示为 $t = \{1, 2, \cdots, T\}$ 。

（2）特征提取

对观测的语音信号来说，可以有很多不同的特征，包括时域和频域的。在语音识别中，常用的方法是利用线性预测编码（LPC）对语音信号进行特征分析。本文先进行 LPC 分析，再将 LPC 系数转化为倒谱系数。记 LPC 分析的阶次为 M，倒谱系数的数目为 Q，为了增加动态信息，将 Q 个倒谱系数的差也作为特征参数，因此特征参数的长度为 $2\times Q$。实际应用时，对每一块的语音信号都进行同样的处理，这样可以得到特征向量序列 $\{y_1, y_2, \cdots, y_T\}$ 。

特征提取的过程可以用下面的函数实现：

```
function y=hmmfeatures(s,N,deltaN,M,Q)
Ns=length(s);                  %信号长度
T=1+fix((Ns-N)/deltaN);   %块的数目
a=zeros(Q,1);
gamma=zeros(Q,1);
gamma_w=zeros(Q,T);
win_gamma=1+(Q/2)*sin(pi/Q*(1:Q)');   %计算倒谱的窗函数
for t=1:T
     idx=(deltaN*(t-1)+1):(deltaN*(t-1)+N);
     sw=s(idx).*hamming(N);
     [rs,eta]=xcorr(sw,M,'biased');
     %基于 Levinson-Durbin 递归的 LPC 分析
     [a(1:M),xi,kappa]=durbin(rs(M+1:2*M+1),M);
     %倒谱系数
     gamma(1)=a(1);
     for i=2:Q
          gamma(i)=a(i)+(1:i-1)*(gamma(1:i-1).*a(i-1:-1:1))/i;
     end
     %加权的倒谱序列
     gamma_w(:,t)=gamma.*win_gamma;
end
%倒谱序列的差
delta_gamma_w=gradient(gamma_w);
%特征向量
y=[gamma_w;delta_gamma_w];
```

（3）矢量量化

为了应用离散概率密度型的 HMM，需要对上述观测的特征向量进行矢量量化，它的作用是产生一个包含 K 个可能的观测向量的码本。这样，通过特征提取过程，从每个单词的一次发音的信号中可以得到观测序列 $\{y_1, y_2, \cdots, y_T\}$ ；再通过矢量量化，产生离散的观测序列 $\{y_1, y_2, \cdots, y_T\}$ 。其中，每个 y_t 可能取 $1 \leqslant k \leqslant K$ 之间的整数（对应码本中的索引）。可以利用



K-均值聚类方法进行矢量量化。

矢量量化的过程可以用如下的函数实现：

```
function [Yc,c,errlog]=kmeans(Y,K,maxiter)
[M,N]=size(Y);
if(K>M)
    error('More centroids than data vectors.')
end
errlog=zeros(maxiter,1);        %每次迭代误差的对数值
%初始聚类中心
perm=randperm(M);
Yc=Y(perm(1:K),:);
d2y=(ones(K,1)*sum((Y.^2)'))';
for i=1:maxiter
    %保留旧聚类中心,以判断是否迭代终止
    Yc_old=Yc;
    %Y 与 Yc 行之间的 Euclidean 距离的平方
    d2=d2y+ones(M,1)*sum((Yc.^2)')-2*Y*Yc';
    %分配 Y 中的每一个向量到最近的中心
    [errvals,c]=min(d2');
    %调整聚类中心
    for k=1:K
        if (sum(c==k)>0)
            Yc(k,:)=sum(Y(c==k,:))/sum(c==k);
        end
    end
    errlog(i)=sum(errvals);
    fprintf(1,'...Iteration %4d...Error %11.6f\n',i,errlog(i));
    %判断终止条件
    if (max(max(abs(Yc-Yc_old)))<10*eps)
        errlog=errlog(1:i);
        return
    end
end
```

（4）模型训练

接下来就可以利用这些码本对 HMM 进行训练，下面以单词“1”的训练为例进行说明，其他单词的训练类似。

其实现的 MATLAB 程序代码如下：

```
>> clear all;
%读取语音信号
load ti46
data=ti46.case(27:36);
L=length(data);
%信号预处理参数
N=320;
```

```
deltaN=80;
M=12;
Q=12;
%矢量量化参数
K=10;
maxiter=500;
%HMM 初始化参数
%状态数
states=5;
%HMM 训练
estA=zeros(5,5,L);
estB=zeros(5,10,L);
%提取特征
for i=1:L
    %初始状态转移概率矩阵
    A0=rand(states,states);
    A0=A0./repmat(sum(A0),states,1);
    B0=rand(K,states);
    B0=(B0./repmat(sum(B0),K,1))';
    for j=1:l
        xdata=load(data{l}{i});
        y=hmmfeatures(xdata,N,deltaN,M,Q);
        %矢量量化
        [yc,c,errlog]=kmeans(y,K,maxiter);
        %训练
        [A0,B0]=hmmtrain(c,A0,B0);
    end
    estA(:,:,i)=A0;
    estB(:,:,i)=B0;
end
```

（5）语音识别

训练完以后，就可以利用这些 HMM 对给定的语音信号进行识别。

```
for i=1:10
    [pStats,logp]=hmmdecode(c,estA(:,:,i),estB(:,:,i));
    p(i)=logp;
end
%概率大小
p
```

概率最大的 HMM 模型对应的单词就是识别的结果。

9.2 示范程序

统计工具箱还提供了一些示范程序，这些程序通过创建一个交互的图形环境来演示统计分布函数、随机数生成器、曲线拟合，以及实验设计函数的用法，见表 9-1。绝大多数示范

程序都提供了图形界面，可以使用自己的数据，而不仅仅是一些例子程序。

表 9-1 示范程序函数

示范程序	目的	示范程序	目的
aoctool	anocova 拟合的交互图形预测	disttool	统计分布的交互图形界面
polytool	多项式拟合的交互图形拟合	randtool	随机生成器的交互控制
robustdemo	鲁棒和最小二乘拟合的交互比较	rsmdemo	实验设计和回归模型



9.2.1 aoctool 演示程序

aoctool 函数

功能：aoctool 是一个通过分析协方差模型来拟合或预测的图形界面。

说明：协方差分析含有一个结果（y，要预测的值）和一个预测数（x，一些用来预测的数据）。通过协方差分析，可以将 y 作为 x 的线性模型，其中数据组之间的协方差可能是互不相同的。Aoctool 函数对以下每组数据进行不同的拟合。

- 相同的均值：$y=\alpha+\varepsilon$。
- 独立的均值：$y=(\alpha+\alpha_i)+\varepsilon$。
- 相同的线上：$y=\alpha+\beta x+\varepsilon$。
- 平行线上：$y=(\alpha+\alpha_i)+\beta x+\varepsilon$。
- 独立的线上：$y=(\alpha+\alpha_i)+(\beta+\beta_i)x+\varepsilon$。

举例来说，在第四个模型中，每个组的截距是不相同的，但是斜率是相同的。在第一个模型中，只有一个常数的截距，没有斜率。为了使方程中的系数容易确定，约定 $\sum\alpha_i=\sum\beta_i=0$。aoctool 函数用 3 个数据窗口来显示拟合结果，第一个窗口显示系数的估计 $(\alpha,\alpha_i,\beta,\beta_i)$。第二个窗口显示了一个变化表，通过这个表，可以决定，一个复杂的模型是否比一个简单的模型更有意义。第三个窗口是主窗口，有以下特征：

- 有重叠拟合线和可选置信区间的数据的图。
- y 轴的文字是用来显示当前 x 值预测的 y 值和 x 值的不确定性，如果当前选中的是一个组。
- 一个数据输入窗口用来计算对一个指定的 x 值的拟合。
- 一个列的窗口用来显示一个指定组的拟合线或显示所有组的拟合线。
- 可以拖动的竖直参考线用来观察变化的 x 值的拟合值。
- 一个关闭按键用于关掉这个演示程序。
- 一个输出列表框用于把拟合后的结果输出到变量中。

【例 9-6】 统计工具箱中有一个很小的数据集合 carsmall，其中包括了一些关于汽车的信息。这很适合用 aoctool 函数来做实验，当然也可以用自己找到的数据。

首先，打开数据库，载入数据。

```
>> load carsmall
who
Your variables are:
```

```
Acceleration   Displacement   MPG       Model_Year   Weight
Cylinders      Horsepower     Model     Origin
```

假设想找汽车的质量和英里数，而且想知道这个关系几年以后是否有效。

第二，打开 aoctool 函数。

```
>> [h,atab,ctab,stats]=aoctool(Weight,MPG,Model_Year)
```

这个工具输出一个主窗口（见图 9-3），一个系数估计窗口（见图 9-4），还有一个变量分析窗口（见图 9-5）。

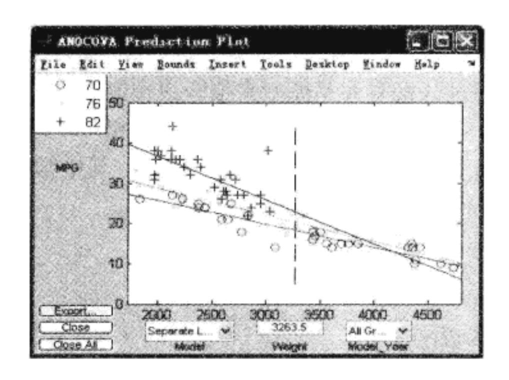


图 9-3 主窗口

每一组的数据都以它自己的符号和颜色表示出来，而且拟合线的颜色与数据组的颜色是相同的。

ANOCOVA Coefficients

File Edit View Insert Tools Desktop Window Help

Coefficient Estimates

Term	Estimate	Std. Err.	T	Prob>|T|
Intercept	45.9798	1.52085	30.23	0
70	-8.5805	1.96186	-4.37	0
76	-3.8902	1.86864	-2.08	0.0403
82	12.4707	2.5568	4.88	0
Slope	-0.0078	0.00056	-14	0
70	0.002	0.00066	2.96	0.0039
76	0.0011	0.00065	1.74	0.0849
82	-0.0031	0.001	-3.1	0.0026

图 9-4 系数估计窗口

最初以线性模型来拟合变量 X，Weight 得到变量 Y，MPG，每一组都有自己独立的直线。这 3 条直线的系数出现在以 ANOCOVA Coefficients 为标题的图形中。可以看到斜率大概都是 0.078，组与组之间仅仅有一点点偏差。

Model Year 70：$y=(45.9798-8.5805)+(-0.0078+0.002)x+\varepsilon$

Model Year 76：$y=(45.9798-3.8902)+(-0.0078+0.0011)x+\varepsilon$

Model Year 82：$y=(45.9798-12.4707)+(-0.0078-0.0031)x+\varepsilon$

可以注意到，这 3 条拟合曲线有几乎一样的斜率，它们实际上相同吗？Model_Year*Weight 表示了斜率的不同，而且 ANOVA 表也对该项的意义做了测试。通过一个 5.23 的 F

统计量和一个 0.072 的 *P* 值，可以知到斜率是明显不相同的。



ANOCOVA Test Results

File Edit View Insert Tools Desktop Window Help

ANOVA Table

Source	d.f.	Sum Sq	Mean Sq	F	Prob>F
Model_Year	2	807.69	403.84	51.98	0
Weight	1	2050.2	2050.2	263.87	0
Model_Year*Weight	2	81.22	40.61	5.23	0.0072
Error	88	683.74	7.77		



图 9-5 ANOCOVA 测试结果

当 3 条线的斜率相同时，为了检验拟合的效果，返回到 ANOCOVA Prediction Plot 窗口（见图 9-3），在【Model】菜单中选择了“Parallel Lines（平行线模型）”，窗口更新返回如图 9-6 所示的结果。

尽管看上去是合理的，但是相对于独立直线模型来说，这是绝对错误的。使用 Model 弹出菜单，恢复到原始状态。

下面介绍 aoctool 函数的两个特性。

- 置信区间。
- 多重比较。

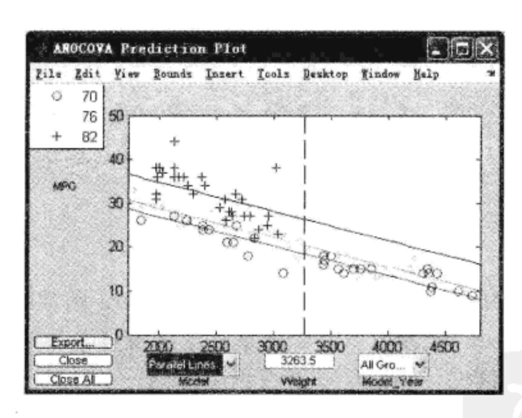


图 9-6 平行线模型窗口

1. 置信区间

已经估计了第一个 Model_Year 的 MPG 和 Weight 的关系，但是它们有多精确呢？如果每次只检查一个组的数据，可以把置信区间和拟合线叠加起来显示。在图形右下角的【Model_Year】菜单中，把设置 All Group 改为 82，其他数据就会消失了，同时置信区间出现在 82 的拟合曲线周围，如图 9-7 所示。

Model_Year 82 的线的周围，用虚线包围了起来。在假设数据满足线性关系的前提下，这个关系为真正的线提供一个 95%的置信区间。注意到拟合其他 Model_Year 时，对于 Weight 在 2000~3000 之间，有一大部分数据落在置信区间的外面。

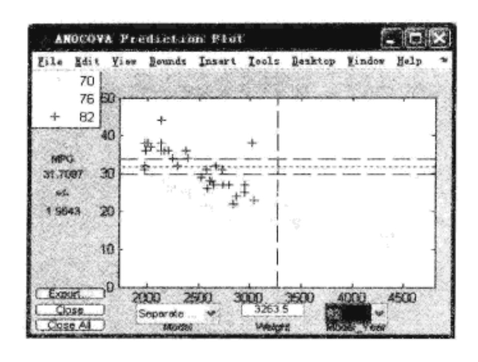


图 9-7　置信区间出现在 82 的拟合曲线周围

有时候，更有意义的是能够对新的观察预测出响应值，不是仅仅估计出平均响应值。aoctool 函数有一个【Bounds】菜单来设置置信区间的定义。用这个菜单把 Line 改为 Observation，如图 9-8 所示，结果区间变宽了，说明了参数估计的不确定性，也说明了新的观察值的随机性。

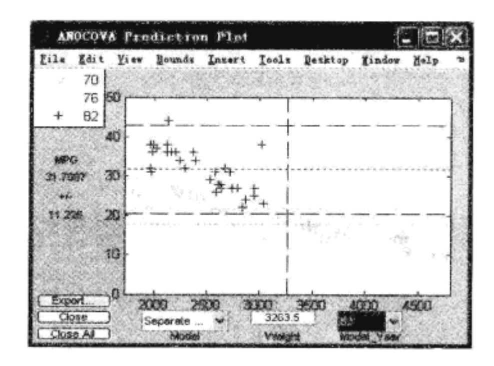


图 9-8　Observation 效果图

2. 多重性

通过把 aoctool 函数的输出结果 stats 作为 multcompare 函数的输入结果，可以做一个多重比较测试。multcompare 函数可以测试斜率、截距或者总体边缘均值。在本例中，已经知道斜率并不是总是一样的，但是能不能有两个相同，而其他的不同呢？下面来检验这个假设。

```
>> multcompare(stats,0.05,'on','','s')
ans =
     1.0000     2.0000    -0.0012     0.0008     0.0029
     1.0000     3.0000     0.0013     0.0051     0.0088
     2.0000     3.0000     0.0005     0.0042     0.0079
```

这个矩阵说明了第一组和第二组（1970 年和 1976 年）的截距的差异是 0.0008，而且这个差异的置信区间是[-0.0012, 0.0029]。这两组之间没有明显的不同，但是 1982 年的截距与其他两个是明显不一样的，图 9-9 显示了相同的信息。

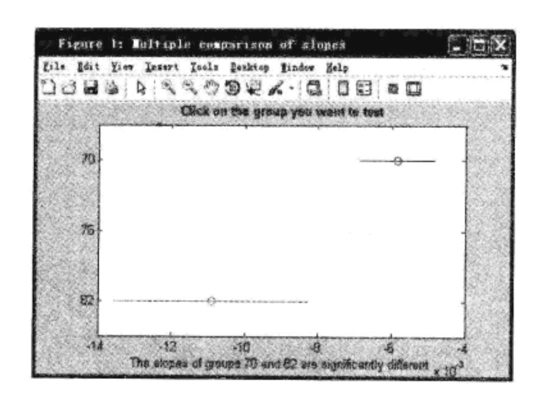


图 9-9　多重比较



9.2.2 disttool 演示程序

disttool 函数

功能：disttool 通过图形环境增加对统计分布的直觉理解。

说明：disttool 示范程序有以下几个特点：

- 对给定的参数 cdf（pdf）画出图形。
- 弹出菜单用来改变分布函数。
- 弹出菜单用来改变分布函数的类型（cdf<->pdf）。
- 滑标用来改变参数设置。
- 数据输入对话框用来选择特别的参数值。
- 数据输入对话框也可以用来限制参数的范围。
- 可以拖动的水平和竖直参考线。
- 一个数据输入窗口可以输入指定的 x 值。
- 对于 cdf 作图，在 y 轴上还有一个数据输入窗口可以查找指定概率的临界值。
- 一个关闭按钮用于结束演示。

其界面如图 9-10 所示。

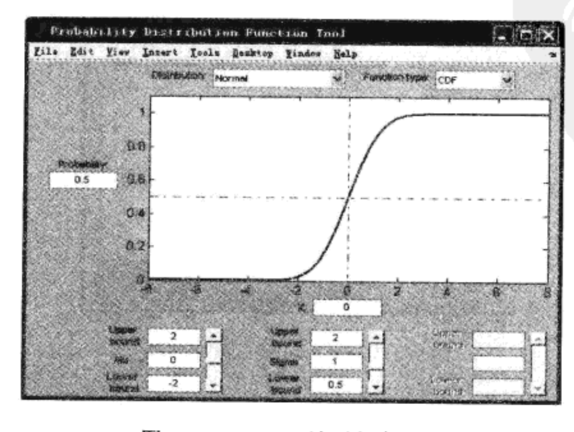


图 9-10　disttool 演示程序界面

9.2.3 polytool 演示程序

polytool 函数

功能：polytool 是多项式曲线拟合的预测图形环境。

说明：polytool 示范程序有以下几个特点：

- 图形区包含数据、拟合的曲线、新预测值的置信区间。
- y 轴上的文字显出 y 的预测值，以及对于当前 x 值的不确定性。
- 数据输入对话框用来改变多项式拟合的程度。
- 数据输入对话框用来计算给定 x 值的多项式的值。
- 可以拖动的水平和竖直参考线。
- 区间和方法菜单可以控制置信区间，以及选择最小二乘拟合还是鲁棒拟合。
- 一个输出列表框可以将拟合后的结果输出到变量中。
- 一个关闭按钮用于结束演示。

用户可以用 polytool 函数对任何数据进行曲线拟合和预测。但是，出于预测的目的，统计工具箱提供了一个数据文件（polydata.dat）来介绍一些基本的概念。

首先，输入这些数据。

```
>> load polydata
>> who
Your variables are:
x    x1   y    y1
```

变量 x 和 y 是从一个 3 次多项式观察（有误差）所得的数据，$x1$ 和 $y1$ 是“真”函数的数据点，没有误差。

如果不指定多项式的次数，polytool 函数只对数据进行线性拟合。

```
>> polytool(x,y)
```

结果如图 9-11 所示。

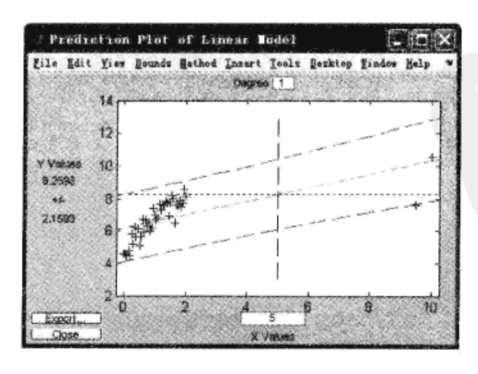


图 9-11　polytool 演示程序界面

线性拟合的效果不是很好，x 值在[0, 2]区间内的大部分的数据比拟合后的曲线坡度还要陡，右边的两个点把拟合的曲线拉了下来。

在顶部的“Degree”文本框，输入“3”，变成 3 次模型。然后，拖动竖直的线到 x=2 的位置，产生如图 9-12 所示的效果。

图 9-12 展示了一个更好的拟合，置信区间更紧地靠在了一起，表示预测中的不确定性更小了。两端的数据在拟合曲线的轨迹上面。



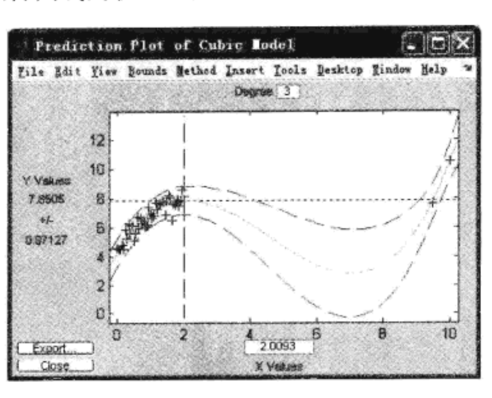


图 9-12 更好的效果

如果这个 3 次多项式拟合得比较好，试一试用更高一点的次数来模拟，看一看精度是不是比较高呢？

因为真正的函数是 3 次的，要用高一点的精度来拟合它，意味可能会过度拟合。在“Degree”文本框中输入“5”，表示用 5 次模型来进行拟合。产生的结果如图 9-13 所示。

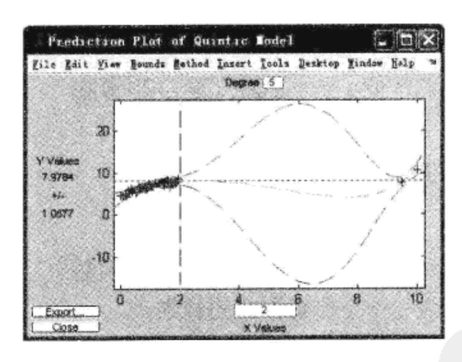


图 9-13 过度拟合

根据置信区间的测量，这个拟合距离数据更加近了。但是在数据之间的区间，预测的不确定性剧烈地增加了。

出现这种置信区间的膨胀现象的主要原因是，数据实际没有包含足够的信息去精确地估计更高次数的多项式，因此即使是插值法在这种情况下，也是有风险的。

9.2.4 randtool 演示程序

randtool 函数

功能：randtool 是一个从多样本的概率中产生随机样本，同时显示直方图的图形界面。

说明：randtool 演示程序有以下特点：

- 一个样本的直方图。
- 一个弹出框用来改变分布函数。
- 滑标用来改变参数设置。
- 一个数据输入对话框用来选择样本大小。
- 数据输入对话框用来选择指定的参数值。
- 数据输入对话框用来选择参数滑标的最大值。
- 一个【Export】按钮用来把当前的样本输出到变量 ans 中。
- 一个【Resample】按钮可以根据固定的样本大小和固定的参数重复地产生随机数。

其界面如图 9-14 所示。

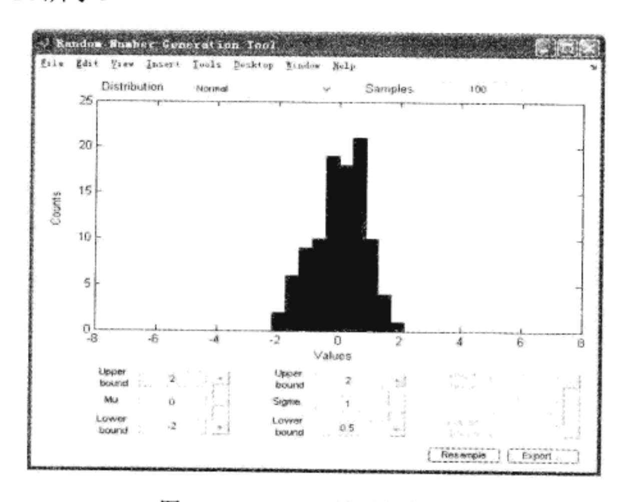


图 9-14 randtool 演示程序界面

9.2.5 robustdemo 演示程序

robustdemo 函数

功能：robustdemo 演示程序是一个图形界面。在该界面中，比较了对一个回应和一个单一的预测数据的最小二乘拟合和一个鲁棒拟合的结果。

说明：要打开这个界面，只需输入函数名即可：

```
>> robustdemo
```

结果图形展现有两条拟合直线的分散点。一条直线是对这些数据的最小二乘回归拟合，另外一条是鲁棒拟合（见图 9-15）。在界面的最底部是每条线的公式和每个拟合的标准偏差的误差估计。

最小二乘拟合的效果主要决定于数据的残留量和各个点的杠杆力量。残留量仅仅是点到直线的竖直距离；杠杆力量主要是衡量每个点距离 X 中心的距离。

鲁棒拟合的效果也依赖于赋于点的权重，点离中心越远，权重越小。

可以用鼠标右键单击每个点，来查看最小二乘拟合的杠杆作用和鲁棒拟合的权重。

在该例中，右边最远的点的杠杆作用为 0.35，它也远离直线，所以它对直线的作用很



大。但是，它有一个很小的权重，因而在鲁棒拟合的时候，它被有效地排除了。

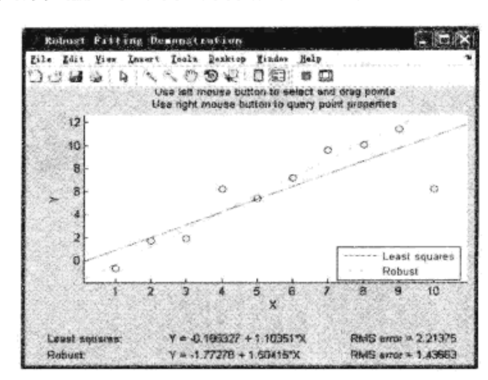


图 9-15　鲁棒拟合图

使用鼠标的左键，可以看到这两条直线是如何变化的。拖动一个点到一个新的地方同时按下鼠标左键不放，当鼠标松开的时候，两条拟合直线都会重绘。

把最右端的点向直线移动，可以看到，两条直线几乎要重合了。现在，这个点几乎占据了所有的鲁棒拟合的权重，如图 9-16 所示。

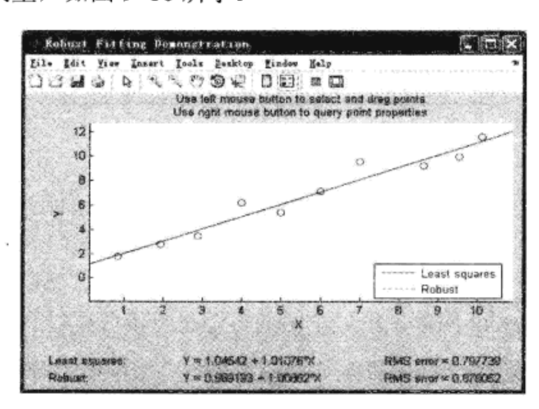


图 9-16　几乎重合的两条直线

9.2.6 rsmdemo 演示程序

rsmdemo 函数

功能：rsmdemo 演示程序是一个通过模拟化学反应来演示实验设计和曲面拟合的图形交互环境。该演示的目的主要是找到反应体的级别以最大化反应率。

说明：该演示程序分为两个部分。

第一部分——比较通过用试错法对实验设计的数据进行搜集。

第二部分——通过非线性模型比较反应曲面模型。

打开这个界面，只需输入函数的名字即可：

```
>> rsmdemo
```

1. 第一部分

开始前，能通过 Reaction Simulator 窗口（见图 9-17）中的滑标来控制 3 种反应体的分压力：氢、n-戊烷和异戊烷。每次单击【Run】按钮，反应体的级别和运行后的结果就进入到 Trial and Error Data 窗口（见图 9-18）中。

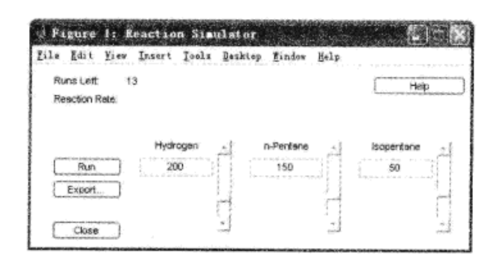


图 9-17　Reaction Simulator 窗口

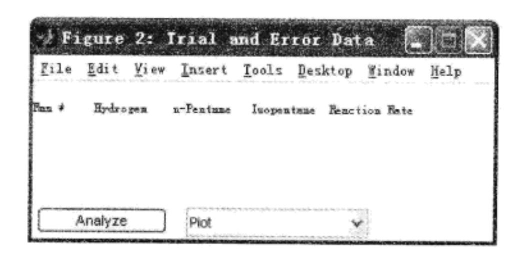


图 9-18　Trial and Error Data 窗口

在前面运行结果的基础上，可以改变反应体的级别来增加反应率。每批只能运行 13 次，当运行到 13 次以后，就可以选择 Trial and Error Data 窗口中的【Plot】菜单或者单击【Analyze】按钮来画出反应体和反应率之间的关系，如图 9-19 所示。当单击 Analyze 按钮时，rsmdemo 函数调用 rstool 函数，用这个函数可以进一步优化得到的效果。

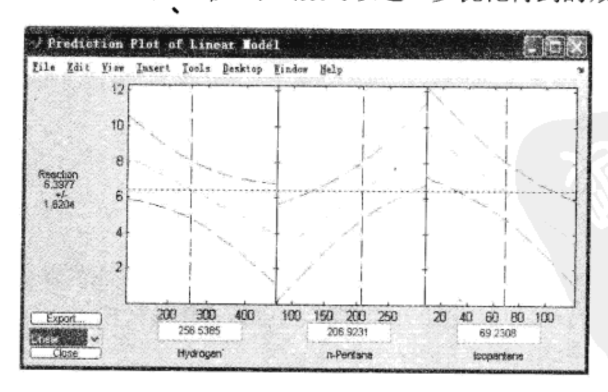


图 9-19　反应体与反应率之间的关系

下一步，从实验设计的结果，执行另外的 13 个运行结果。在 Experimental Data 窗口（见图 9-20）中，单击【Do Experiment】按钮，rsmdemo 函数调用 cordexch 函数产生一个 D-最优设计，然后对每一次运行产生一次反应率（见图 9-21）。

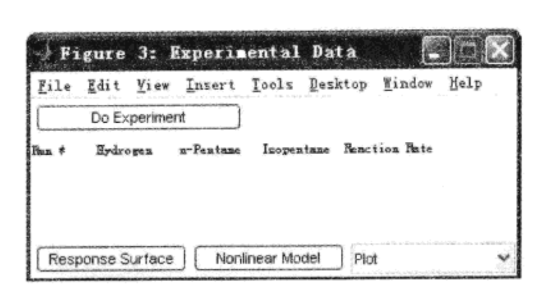




图 9-20 实验数据

现在使用 Experimental Data 窗口中的【Plot】菜单画出反应体与反应率之间的关系，或者单击【Response Surface】按钮调用 rstool 函数发现反应体的最优级别。比较前面生成的两个结果，可能会发现下面的一些不同之处：

- 可以通过实验设计得到的数据拟合一个二次的模型，但是通过试错法得到的数据可能就不能满足一个二次模型或是交互作用模型。
- 使用实验设计得到的数据，可以更容易地发现反应体的级别，从而最大化反应率。甚至假设发现试错法中的最好的设置，那么置信区间也会比从实验设计中得到的置信区间更宽。

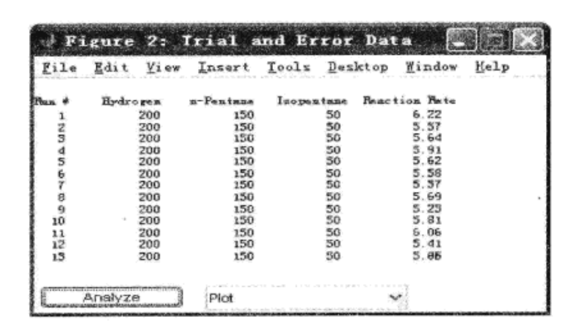


图 9-21 产生实验数据显示

2. 第二部分

现在通过多项式模型和非线性模型分析用实验设计得到的数据，并比较它们的结果。用来生成数据的真正的过程，实际上是一个真正的模型。在数据的范围内，一个二次模型可以很好地近似于这个真正的模型。

对于多项式模拟，单击 Experimental Data 窗口中的【Response Surface】按钮。Rsmdemo 函数调用 rstool 函数，rstool 函数用一个二次模型来拟合这些数据。拖动参考线改变反应体的级别，同时发现最优的反应率。观察置信区间的宽度。

现在单击【Nonlinear Model】按钮，rsmodemo 函数调用 nlintool 函数。nlintool 函数用 Hougen-Watson 模型来拟合这些数据，如图 9-22 所示。

和二次模型一样，可以拖动参考线改变当前反应体的级别，观察反应率和置信区间。

比较刚刚得到的两组结果，即使这个真正的模型是非线性的，也会发现多项式模型提供了一个很好的拟合。因为多项式比非线性拟合处理起来更容易，所以多项式模型常常是更可

取的，即使原来的模型是非线性的。但是，有一点值得注意，多项式模型用来推测数据区域外的数据是不可靠的。

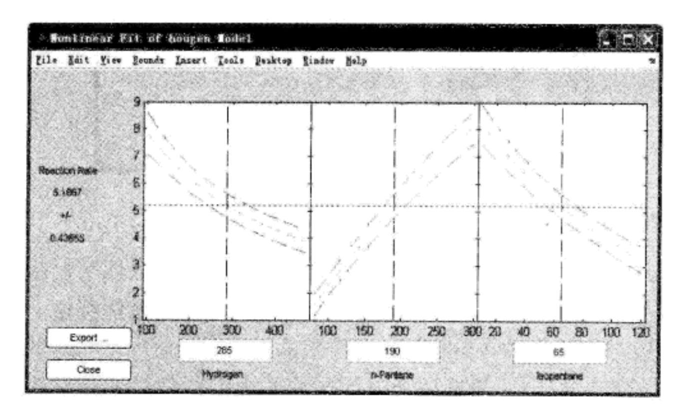
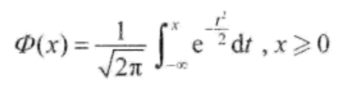


图 9-22　Hougen-Watson 模拟拟合

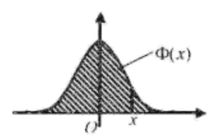


附　　录

附录 A　标准正态分布函数表



$$\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}\mathrm{e}^{-\frac{t^2}{2}}\mathrm{d}t\ ,\ x\geqslant 0$$



x	0.00	0.01	0.02	0.03	0.04
0.0	0.5000	0.5040	0.5080	0.5120	0.5160
0.1	0.5398	0.5438	0.5478	0.5517	0.5557
0.2	0.5793	0.5832	0.5871	0.5910	0.5948
0.3	0.6179	0.6217	0.6255	0.6293	0.6331
0.4	0.6554	0.6591	0.6628	0.6664	0.6700
0.5	0.6915	0.6950	0.6985	0.7019	0.7054
0.6	0.7257	0.7291	0.7324	0.7357	0.7389
0.7	0.7580	0.7611	0.7642	0.7673	0.7703
0.8	0.7881	0.7910	0.7939	0.7962	0.7995
0.9	0.8159	0.8186	0.8212	0.8238	0.8264
1.0	0.8413	0.8438	0.8461	0.8485	0.8508
1.1	0.8643	0.8665	0.8686	0.8708	0.08729
1.2	0.8849	0.8869	0.8888	0.8907	0.8925
1.3	0.90320	0.90490	0.90678	0.90824	0.90988
1.4	0.91924	0.92073	0.92220	0.92364	0.92507
1.5	0.93319	0.93448	0.93574	0.93699	0.93882
1.6	0.94520	0.94630	0.94738	0.94845	0.94950
1.7	0.95543	0.95637	0.95728	0.95818	0.95907
1.8	0.96407	0.96485	0.96562	0.96638	0.96712
1.9	0.97128	0.97193	0.97257	0.97320	0.97381
2.0	0.97725	0.97778	0.97831	0.97882	0.97932
2.1	0.98214	0.98257	0.98300	0.98341	0.98382
2.2	0.98610	0.98645	0.98679	0.98713	0.98745
2.3	0.98928	0.98956	0.98983	0.99010	0.99036
2.4	0.99180	0.99205	0.99224	0.99245	0.99266
2.5	0.99379	0.99396	0.99413	0.99430	0.99446
2.6	0.99534	0.99547	0.99560	0.99573	0.99586
2.7	0.99653	0.99664	0.99674	0.99683	0.99693
2.8	0.99745	0.99752	0.99760	0.99767	0.99774
2.9	0.99813	0.99819	0.99825	0.99831	09836
3.0	99865	0.99869	0.99874	0.99878	0.99982
3.1	0099903	0.99906	0.99910	0.99913	0.99916
3.2	0.99931	0.99934	0.99936	0.99938	0.99940
3.3	0.99952	0.99953	0.99955	0.99957	0.99958
3.4	0.99966	0.99968	0.99969	0.99970	0.99971

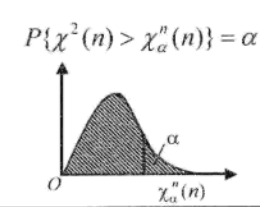
（续）

x	0.00	0.01	0.02	0.03	0.04
3.5	0.99977	0.99978	0.99978	0.99979	0.99980
3.6	0.99984	0.99985	0.99985	0.99986	0.99986
3.7	0.99989	0.99990	0.99990	0.99990	0.99991
3.8	0.99993	0.99993	0.99993	0.99994	0.99994
3.9	0.99995	0.99995	0.99996	0.99996	0.99996
4.0	0.99997	0.99997	0.99997	0.99997	0.99997
4.1	0.99998	0.99998	0.99998	0.99998	0.99998
4.2	0.99999	0.99999	0.99999	0.99999	0.99999
4.3	0.99999	0.99999	0.99999	0.99999	0.99999
4.4	0.99999	0.99999	1.00000	1.00000	1.00000
x	0.05	0.06	0.07	0.08	0.09
0.0	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6736	0.6772	0.6808	0.6443	0.6879
0.5	0.7088	0.7123	0.7157	0.6808	0.7224
0.6	0.7422	0.7454	0.7486	0.7157	0.7549
0.7	0.7734	0.7764	0.7794	0.7486	0.7852
0.8	0.8023	0.8051	0.8078	0.7794	0.8133
0.9	0.8298	0.8315	0.8340	0.8078	0.8389
1.0	0.8531	0.8554	0.8577	0.8340	0.8621
1.1	0.8749	0.8770	0.8790	0.8577	0.8830
1.2	0.8944	0.8962	0.8980	0.8790	0.90147
1.3	0.91140	0.91309	0.91466	0.8980	0.91774
1.4	0.92647	0.92785	0.92922	0.91621	0.93189
1.5	0.93943	0.94062	0.94179	0.993056	0.94408
1.6	0.95053	0.95154	0.95254	0.94296	0.95449
1.7	0.95994	0.96080	0.96164	0.95352	0.96327
1.8	0.96784	0.96856	0.96926	0.96246	0.97062
1.9	0.97441	0.97500	0.97558	0.96995	0.97670
2.0	0.97982	0.98030	0.98077	0.97615	0.98169
2.1	0.98422	0.98461	0.98500	0.98124	0.98574
2.2	0.98778	0.98809	0.98840	0.98537	0.98899
2.3	0.99061	0.99086	0.99111	0.98870	0.99158
2.4	0.99286	0.99305	0.99324	0.99134	0.99361
2.5	0.99461	0.99477	0.99492	0.99343	0.99520
2.6	0.99598	0.99609	0.99621	0.99506	0.99643
2.7	0.99702	0.99711	0.99720	0.99632	0.99737
2.8	0.99781	0.99788	0.9979	0.99728	0.99807
2.9	0.99841	0.99846	599851	0.99801	0.99861
3.0	0.99886	0.99889	0.99893	0.99856	0.99900
3.1	0.99918	0.99921	0.99924	0.99897	0.99929
3.2	0.99942	0.999944	0.99946	0.99948	0.99950
3.3	0.99960	0.99961	0.99962	0.99964	0.99965
3.4	0.99972	0.99973	0.99974	0.99975	0.99976
3.5	0.99981	0.99981	0.99982	0.99983	0.99983
3.6	0.99987	0.99987	0.99988	0.99988	0.99989
3.7	0.99991	0.99992	0.99992	0.99992	0.99992
3.8	0.9994	0.99994	0.99995	0.99995	0.99995
3.9	0.99996	0.99996	0.99996	0.99997	0.99997
4.0	0.99997	0.99998	0.99998	0.99998	0.99998
4.1	0.99998	0.99998	0.99998	0.99999	0.99999
4.2	0.99999	0.99999	0.99999	0.99999	0.99999
4.3	0.99999	0.99999	0.99999	0.99999	0.99999
4.4	1.00000	1.00000	1.00000	1.00000	1.00000



附录 B　χ^2 分布上侧分位点表

$$P\{\chi^2(n) > \chi_\alpha^n(n)\} = \alpha$$



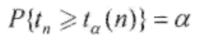
α / n	$\alpha = 0.995$	$\alpha = 0.99$	$\alpha = 0.975$	$\alpha = 0.95$	$\alpha = 0.90$	$\alpha = 0.75$
1	—	—	0.001	0.004	0.016	0.102
2	0.010	0.202	0.051	0.103	0.211	0.575
3	0.072	0.115	0.216	0.352	0.584	1.213
4	0.207	0.297	0.484	0.711	1.064	1.923
5	0.412	0.554	0.831	1.145	1.610	2.675
6	0.676	0.872	1.237	1.635	20204	3.455
7	0.989	1.239	1.690	2.167	2.833	4.255
8	1.344	1.646	2.180	2.733	3.490	5.071
9	1.735	2.088	2.700	3.325	4.168	5.899
10	2.156	2.558	3.247	3.940	4.865	6.737
11	2.603	3.053	3.816	4.575	5.578	7.584
12	3.074	3.571	4.404	5.5226	6.304	8.438
13	3.565	4.107	5.009	5.892	7.042	9.299
14	4.075	4.660	5.629	6.571	7.790	10.165
15	4.601	5.229	6.262	7.261	8.547	11.037
16	5.142	5.812	6.908	7.962	9.312	11.912
17	5.697	6.408	7.564	8.672	10.085	12.792
18	6.265	7.015	8.231	9.390	10.865	13.675
19	6.844	7.633	8.907	10.117	11.651	14.562
20	7.434	8.260	9.591	10851	12.443	15.452
21	8.034	8.897	10.283	11.591	13.240	16.344
22	8.643	9.542	10.982	12.338	14.042	17.240
23	9.260	10.196	11.689	13.091	14.848	18.137
24	9.886	10.856	12.401	13.848	15.659	19.037
25	10.520	11.524	13.120	14.611	16.473	19.939
26	11.160	12.198	13.844	15.379	17.292	20.843
27	11.808	12.879	14.573	16.151	18.114	21.749
28	12.461	13.565	15.308	16.928	18.939	22.657
29	13.121	14.257	16.047	17.708	19.768	23.567
30	13.787	14.954	16.791	18.493	20.599	24.478
31	14.458	15.655	17.539	19.281	21.434	25.390
32	15.134	16.362	18.291	20.072	22.271	26.304
33	15.815	17.074	19.047	20.867	23.110	27.219
34	16.501	17.789	19.806	21.664	23.952	28.136
35	17.192	18.509	20.569	22.456	24.797	29.054
36	17.887	19.233	21.336	23.269	25.643	29.973
37	18.586	19.960	22.106	24.075	26.492	30.893
38	19.289	20.691	22.878	24.884	27.343	31.815
39	19.996	21.426	23.654	25.695	28.196	32.737
40	20.707	22.164	24.433	26.509	29.051	33.660
41	21.421	22.906	25.215	27.326	29.907	34.585
42	22.138	23.650	25.999	28.144	30.765	35.510
43	22.859	24.398	26.785	28.965	31.325	36.369
44	23.584	25.148	27.575	29.787	32.487	37.363
45	24.311	25.901	28.366	30.612	33.350	38.291



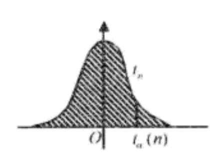
（续）

n \ α	$\alpha=0.25$	$\alpha=0.10$	$\alpha=0.05$	$\alpha=0.025$	$\alpha=0.01$	$\alpha=0.005$
1	1.323	2.706	3.841	5.024	6.635	7.879
2	2.773	4.605	5.991	7.378	9.210	10.597
3	4.108	6.251	7.815	9.348	11.345	12.833
4	5.385	7.779	9.488	11.143	13.277	14.860
5	6.626	9.236	11.071	12.833	15.086	16.750
6	7.841	10.450	12.592	14.449	16.812	18.548
7	9.037	12.017	14.067	16.013	18.475	20.278
8	10.219	13.362	15.507	17.535	20.090	21.955
9	11.389	14.684	16.919	19.023	21.666	23.589
10	12.549	15.987	18.307	20.483	23.209	25.188
11	13.701	17.275	19.675	21.920	24.725	26.756
12	14.845	18.549	21.026	23.337	26.217	28.299
13	15.984	19.812	22.362	24.736	27.688	29.819
14	17.117	21.064	23.685	26.119	29.141	31.319
15	18.245	22.307	24.996	27.488	30.578	32.801
16	19.369	23.542	26.296	28.845	32.000	34.267
17	20.489	24.769	27.587	30.191	33.409	35.718
18	21.605	25.989	28.869	31.526	34.805	37.156
19	22.718	27.204	30.144	32.852	36.191	38.582
20	23.828	28.412	31.410	34.170	37.566	39.997
21	24.935	29.615	32.671	35.479	38.932	41.401
22	26.039	30.813	33.924	36.781	40.289	42.796
23	27.141	32.007	35.172	38.076	41.638	44.181
24	28.241	33.196	36.415	39.364	42.980	45.559
25	29. 339	34.382	37.625	40.646	44.314	46.928
26	30.435	35.563	38.885	41.923	45.642	48.290
27	31.528	36.741	40.113	43.194	46.963	49.645
28	32.620	37.916	41.337	44.461	48.278	50.993
29	33.711	39.087	42.557	45.722	49.588	52.336
30	34.800	40.256	43.773	46.979	50.892	53.672
31	35.887	41.422	44.985	48.232	52.191	55.003
32	36.973	42.585	46.194	49.480	53.486	56.328
33	38.058	4.3745	47.400	50.725	54.776	57.648
34	39.141	44.903	48.602	51.966	56.061	58.964
35	40.223	46.059	49.802	53.203	57. 342	60.275
36	41.304	47.212	50.998	54.437	58.619	61.581
37	42.383	48.363	52.192	55.668	59.892	62.883
38	43.462	49.513	53.384	56.896	61.162	64.181
39	44.539	50.660	54.572	58.120	62.428	65.476
40	45.616	51.805	55.758	59.342	63.691	66.766
41	46.692	52.949	56.942	60.561	64.950	68.053
42	47.766	54.090	58.124	61.777	66.206	69.336
43	48.840	55.230	59.304	62.990	67.459	70.616
44	49.913	56.369	60.481	64.201	68.710	71.893
45	50.985	57.505	61.656	65.410	69.957	73.166

附录 C　t 分布上侧分位点表



$$P\{t_n \geqslant t_\alpha(n)\} = \alpha$$





n \ α	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.01$	$\alpha = 0.005$
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.356	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
40	1.303	1.684	2.021	2.423	2.704
60	1.296	1.671	2.000	2.390	2.660
120	1.289	1.658	1.980	2.358	2.617
∞	1.282	1.645	1.960	2.326	2.576

参 考 文 献

[1] 王明慈，沈恒范. 概率论与数理统计[M]. 北京：高等教育出版社，1999.

[2] 贺子兴，等. 概率论与数理统计[M]. 北京：科学出版社，2000.

[3] 刘卫国. 科学计算与 MATLAB 语言[M]. 北京：中国铁道出版社，2000.

[4] 王松桂，等. 概率论与数理统计[M]. 北京：科学出版社，2002.

[5] 于义良，张银生. 实用概率统计[M]. 北京：中国人民大学出版社，2002.

[6] 梅常林，周家良. 实用统计方法[M]. 北京：科学出版社，2002.

[7] 何强，何英. MATLAB 扩展编程[M]. 北京：清华大学出版社，2002.

[8] 奥特 R L，朗格内克 M. 统计学方法与数据分析引论[M]. 北京：科学出版社，2003.

[9] 章昕. 概率统计双博士课堂[M]. 北京：机械工业出版社，2003.

[10] 丁杰，高文杰. 概率统计[M]. 天津：天津大学出版社，2004.

[11] 汤大林，等. 概率论与数理统计[M]. 天津：天津大学出版社，2004.

[12] 薛定宇，陈阳泉. 高等应用数学问题的 MATLAB 求解[M]. 北京：清华大学出版社，2004.

[13] 苏金明，王永利. MATLAB 7.0 实用指南：上册[M]. 北京：电子工业出版社，2004.

[14] 姜启源，邢文训，等. 大学数学实验[M]. 北京：清华大学出版社，2005.

[15] 陈仲生. 基于 MATLAB 7.0 的统计信息处理[M]. 长沙：湖南科学技术出版社，2005.

[16] 杨振海，张忠占. 应用数理统计[M]. 北京：北京工业大学出版社，2005.

[17] 沈邦兴，文昌俊. 实验设计与工程应用[M]. 北京：中国计量出版社，2005.

[18] 高惠旋. 应用多元统计分析[M]. 北京：北京大学出版社，2005.

[19] 王岩，隋思涟，王家青. 数理统计与 MATLAB 工程数据分析[M]. 北京：清华大学出版社，2006.

[20] 李勇，张淑敏. 统计学导论[M]. 北京：人民邮电出版社，2006.

[21] 庄楚强，何春雄. 应用数理统计基础[M]. 广州：华南理工大学出版社，2006.

[22] 吴礼斌，李柏年. 数学实验与建模[M]. 北京：国防工业出版社，2007.

[23] 刘金兰. 管理统计学[M]. 天津：天津大学出版社，2007.

[24] 许国根，许萍萍. 化学化工中的数学方法及 MATLAB 实现[M]. 北京：化学工业出版社，2008.

[25] 包研科，李娜. 数理统计与 MATLAB 数据处理[M]. 沈阳：东北大学出版社，2008.

[G e n e r a l I n f o r m a t i o n]
书名＝M A T L A B概率与数理统计分析
S S号＝1 2 4 4 6 3 7 6